BUMP: A Benchmark of Unfaithful Minimal Pairs for Meta-Evaluation of Faithfulness Metrics

Liang Ma[1] **Shuyang Cao**[2] **Robert L. Logan IV**[1] **Di Lu**[1]
Shihao Ran[1] **Ke Zhang**[1] **Joel Tetreault**[1] **Alejandro Jaimes**[1]
[1]Dataminr Inc. [2]University of Michigan, Ann Arbor
{lma,rlogan,dlu,sran,kzhang,jtetreault,
ajaimes}@dataminr.com caoshuy@umich.edu

Abstract

The proliferation of automatic faithfulness metrics for summarization has produced a need for benchmarks to evaluate them. While existing benchmarks measure the correlation with human judgements of faithfulness on model-generated summaries, they are insufficient for diagnosing whether metrics are: 1) *consistent*, i.e., indicate lower faithfulness as errors are introduced into a summary, 2) effective on *human-written* texts, and 3) sensitive to different *error types* (as summaries can contain multiple errors). To address these needs, we present a benchmark of unfaithful minimal pairs (BUMP), a dataset of 889 *human-written*, *minimally different* summary pairs, where a single error is introduced to a summary from the CNN/DailyMail dataset to produce an unfaithful summary. We find BUMP complements existing benchmarks in a number of ways: 1) the summaries in BUMP are harder to discriminate and less probable under SOTA summarization models, 2) unlike non-pair-based datasets, BUMP can be used to measure the consistency of metrics, and reveals that the most discriminative metrics tend not to be the most consistent, and 3) unlike datasets containing generated summaries with multiple errors, BUMP enables the measurement of metrics' performance on individual error types.

1 Introduction

Although modern abstractive summarization systems have improved in their ability to produce fluent text (Lewis et al., 2020), their ability to generate text that is factually grounded in the source article remains an issue (Kryscinski et al., 2020). This phenomenon has inspired the NLP community to develop faithfulness evaluation metrics (Fabbri et al., 2022; Laban et al., 2022; Honovich et al., 2021; Scialom et al., 2021) that automatically measure the extent to which abstractive summarization systems produce summaries that contain information that cannot be verified by the source article.

As the number of these automatic faithfulness metrics has increased, there has arisen a corresponding need for benchmarks that evaluate their relative strengths. To satisfy this need, researchers have developed datasets such as FRANK (Pagnoni et al., 2021) and TRUE (Honovich et al., 2022) that are comprised of model-generated summaries along with human-annotated faithfulness levels. Although these datasets are useful for evaluating the degree to which faithfulness metrics correlate with human judgements and can discriminate unfaithful summaries, a number of factors limit the conclusions that can be drawn from them. For one, because model summaries can vary in terms of length, content, and number of errors, these benchmarks are ill-suited for drawing conclusions about the *consistency* (Gabriel et al., 2021) of metrics, i.e., whether their scores indicate lower faithfulness as summaries become increasingly unfaithful, as well as their sensitivity to specific *types of errors* since summaries can contain multiple errors. Furthermore, as the summaries are machine-generated, these benchmarks cannot evaluate whether metrics can detect *human-written* unfaithful summaries.

To enable research on these topics, we present BUMP—a benchmark of unfaithful minimal pairs—a dataset of 889 minimally different summary pairs where all unfaithful summaries are generated by *human annotators*. As illustrated in Figure 1, given an article and its reference summary, we ask a human annotator to edit the reference summary in a minimal way such that the edited summary exhibits one unfaithful error. We design two tasks for performance comparisons: 1) taxonomy-based edits, where a specific unfaithfulness error type is required according to our proposed taxonomy, and 2) freestyle edits, where no error type constraints are imposed. The motivation behind the first task setting is to ensure that different error types are adequately represented in our dataset, while the second task setting is important for understanding

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12788–12812
July 9-14, 2023 ©2023 Association for Computational Linguistics

Article: Manchester City and Chelsea are set to battle it out for the signature of West Ham left-back Aaron Cresswell this summer... West Ham snapped up Cresswell for £2million from Ipswich last summer but he has adapted to the top tier with relative ease and attracted the eye of the division's Champions League clubs... England stars Jack Wilshere, Jordan Henderson and Liverpool contract rebel Raheem Sterling are also on City's wanted list to fulfil the quota. Premier League sides must name eight homegrown players in their 25-man squad registered at the beginning of the season. A player is deemed homegrown if they have spent at least three years at any English club before the age of 21, regardless of nationality.

Summary [Extrinsic Entity Error]: Aaron Cresswell has impressed during debut season in Premier League . The left back joined West Ham from Championship club Ipswich for £2m . Manchester City and Chelsea are both keen to sign the 25-year-old . Both clubs are mindful of boosting their quota of ~~homegrown~~ foreign players .

Faithfulness Metric Scores

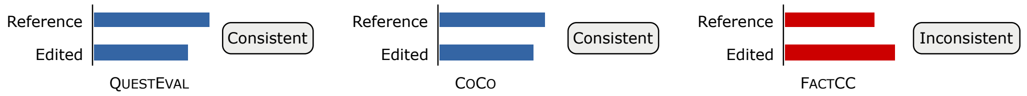


Figure 1: **Example from BUMP dataset.** An annotator constructs an unfaithful summary containing an *Extrinsic Entity Error* (Section 3.2) by replacing the word "homegrown" in the reference summary with the word "foreign". The reference and edited summary form a minimal unfaithful summary pair. Faithfulness metrics are evaluated on both the reference and edited summary and compared to measure whether the metric is consistent, e.g., in this example, QUESTEVAL and COCO are consistent, while FACTCC is not.

the completeness of our error type taxonomy as well as whether annotation difficulty is affected by instructing annotators to focus on specific error types.

We use BUMP to study the ability and performance consistency of faithfulness evaluation metrics in differentiating unfaithful summaries from faithful ones. Similar to how minimal pairs are used to diagnose linguistic knowledge of language models (Marvin and Linzen, 2018; Warstadt et al., 2020), the minimal summary pairs in BUMP allow targeted tests of a metric's consistency on different types of errors (Table 1). This setup minimizes the effect of confounding factors that affect similar analyses (e.g., Pagnoni et al. (2021) and Tang et al. (2022)) such as text length, stylistic variation, and multiple errors occurring in the same summary. We evaluate standard and state-of-the-art faithfulness metrics on BUMP using meta-evaluation protocols that target two phenomena: 1) *consistency*, i.e. the fraction of unfaithful summaries that receive a lower score than their corresponding faithful summaries, and 2) *discriminability*, i.e., the metric's ability to classify unfaithful vs. faithful summaries as measured by ROC AUC.

Our results (Section 5) yield a number of useful findings: 1) *BUMP differs substantially from existing benchmarks*: the summaries in BUMP are harder to discriminate (ROC AUC scores between 50–70% vs. 70–84%) and are less probable under SOTA summarization models. 2) *Discriminability != consistency*: interestingly, the most consistent metrics (BARTSCORE, COCO) tend to have poor discriminability. 3) *Some error types are harder than others*: e.g., metrics seem to uniformly struggle with summaries containing *Intrinsic Error*s.

In sum, our contributions are three-fold: 1) We build a benchmark of human-generated unfaithful minimal pairs (BUMP) for evaluating faithfulness metrics. 2) We show human-generated unfaithful errors are substantially different from and more challenging than model-generated ones. 3) We demonstrate how BUMP provides insights on both the consistency and discriminative ability of faithfulness metrics on different error types than prior evaluation benchmarks that complement insights from existing benchmarks. The BUMP dataset is available at: `https://github.com/dataminr-ai/BUMP`.

2 Related Work

Standard evaluation metrics for text generation tasks, e.g., BLEU and ROUGE, do not correlate well with human judgements of factual alignment in summarization settings (Kryscinski et al., 2019; Maynez et al., 2020). This has motivated the development of automated faithfulness metrics that quantify factual alignment through methods that either: use NLI to measure the entailment degree between the source article and summary (Kryscinski et al., 2020; Goyal and Durrett, 2020; Laban et al., 2022), compare summary probabilities when relevant information is removed from the source (Xie et al., 2021), or use question answering models to measure if questions derived from the source can be answered by the summary and vice versa (Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021).

Existing faithfulness metric evaluations use one of two classes of benchmarks: 1) machine-generated summaries paired with human-annotated faithfulness levels (Laban et al., 2022; Pagnoni et al., 2021; Tang et al., 2022), and 2) summary pairs pertaining to the same article where one summary is faithful and the other is unfaithful (Falke et al., 2019; Gabriel et al., 2021). While both classes can evaluate a metric's ability to discriminate unfaithful summaries, the latter additionally allows for consistency tests, i.e., whether metrics assign higher values to more faithful summaries.

The BUMP dataset belongs to the second class of benchmarks; however, it has a number of unique properties. First, unlike both Falke et al. (2019) and Gabriel et al. (2021), the unfaithful summaries in BUMP are human-written. In addition, the unfaithful summaries in BUMP are *minimally different*, in the sense that only a single error differentiates the faithful and unfaithful summary. As shown in Section 5, this produces summary pairs that are substantially more challenging for metrics to differentiate. Inspired by the use of minimal pairs to diagnose linguistic knowledge of language models (Marvin and Linzen, 2018; Warstadt et al., 2020), the benefit of this approach is that it allows targeted tests of a metric's consistency on different types of errors (Section 3.2) while minimizing the effect of confounding factors. Therefore, unlike other benchmarks with error type annotations (Pagnoni et al., 2021; Tang et al., 2022), results on BUMP are not complicated by issues such as multiple errors appearing in the same summary.

3 Benchmark of Unfaithful Minimal Pairs (BUMP)

Two annotation tasks are designed for BUMP, where Task 1 is taxonomy-based (a specific error type is required for the edited summary), and Task 2 allows freestyle edits (i.e., no error type constraints are imposed). In this section, we first describe how data sources are selected to build BUMP (3.1), and then describe the details of the two annotation tasks (3.2 and 3.3).

3.1 Dataset

For Task 1, we randomly select 100 article-summary pairs from the test set of the CNN/DailyMail dataset (Hermann et al., 2015).[1] For Task 2, we select an additional 100 random article-summary pairs. Both tasks are performed via Amazon Mechanical Turk.[2]

3.2 Task 1: Taxonomy-based Unfaithful Summaries

To obtain fine-grained performance evaluations of faithfulness metrics, it is critical to evaluate their sensitivity regarding various error types. Furthermore, benchmarks should contain sufficiently many instances associated with each error type to enable statistically significant comparisons to be made. To this end, we first define a taxonomy of unfaithful error types, and then ask annotators to introduce errors of a specific type in order to ensure each error type is adequately represented in the final dataset.

We note that existing taxonomies of error types may contain overlapped error types, e.g., grammatical vs. entity errors in FRANK (Pagnoni et al., 2021) or lack fine granularity, e.g., Tang et al. (2022). By considering the strengths and shortcomings of existing taxonomies, we define our own taxonomy in Table 1. Our taxonomy is first adapted from the one in FRANK (Pagnoni et al., 2021) by including semantic frame errors (*Predicate Error*, *Entity Error*, and *Circumstance Error*) and *Coreference Error*, and removing error types that might overlap with others. To further categorize each semantic frame error, we adopt the notions of *Intrinsic* and *Extrinsic* errors (Maynez et al., 2020; Goyal and Durrett, 2020; Tang et al., 2022). Note that we do not simply categorize errors into the *Intrinsic* and *Extrinsic* ones, as we believe semantic frame errors can better instruct annotators to create summaries with diverse unfaithful errors. In our taxonomy, the *Intrinsic/Extrinsic* distinction only applies to the *Predicate, Entity, and Circumstance Error*, since for a *Coreference Error*, it is generally ambiguous whether an erroneous pronoun/reference that does not exist in the source article should be regarded as intrinsic or extrinsic. In total, this results in seven different error types.

[1] We do not annotate samples from the XSum dataset (Narayan et al., 2018) since the reference summaries are frequently unfaithful (Maynez et al., 2020).

[2] `https://www.mturk.com/`; annotation guidelines and interfaces are detailed in Appendices A and B.

For each of the seven error types in this taxonomy, given an article-summary pair, we ask the annotator to introduce an error of the required type through a minimal edit to the reference summary. All <*article, summary, error type*> Human Intelligence Tasks (HITs) in Amazon Mechanical Turk are shuffled and there is no annotation repetition, i.e., one assignment per HIT. This increases the chance that edits of the same reference summary will be made by different annotators. Additional details regarding qualification tests and annotation instructions are presented in Appendix A.

After the data collection, we manually check the validity of each edit. For cases where the edits do not match the required error types, we relabel them with the corrected error types based on our taxonomy. The dataset statistics after correction are shown in Table 2. For this task, one common mistake is that annotators consider the quantity of a *noun object* as a circumstance and make edits to the quantity (the first example in Table 3), hence mistakenly treat *Entity Error*s as *Circumstance Error*s, which causes the total number of *Circumstance Error*s to be only 160 (much smaller than that of *Entity Error*s; see Table 2). Another frequent mistake is that the edited word actually exists in the original article for the required extrinsic error (the second example in Table 3), which results in a smaller number of *Extrinsic Error*s than intrinsic ones across all semantic frame errors, especially for *Predicate Error*s. Furthermore, Table 2 shows all edited summaries can be categorized by our taxonomy (no summaries are relabeled as "Other"), and the incorrect response rate is 16%, suggesting that, in general, annotators correctly respond with the required error types.

3.3 Task 2: Freestyle Unfaithful Summaries

In addition to the taxonomy-based Task 1, we also conduct a separate task, Task 2, where annotators can edit reference summaries in any way they want, i.e., freestyle editing, as long as only one error is introduced to the reference summary via minimal edits. The goal of Task 2 is to understand how human-generated unfaithful summaries may vary, and how the performance of faithfulness evaluation metrics changes accordingly, when there are no error type constraints. In particular, only annotators who did not participate in the qualification test of Task 1 are considered to participate in this task; in this way, we ensure the edited summaries in Task 2 are not constrained to any known error types.

Error Type	Description
Predicate Error	The predicate in the summary is inconsistent with the source article.
Entity Error	The subject/object of a predicate is inconsistent with the source article.
Circumstance Error	Time, duration, or location of an event of the predicate is wrong.
Coreference Error	A pronoun/reference with wrong or nonexistent antecedent.
Intrinsic Error	Error derived from information within the source article.
Extrinsic Error	Error contains information not present in the source article.

Table 1: **Error type taxonomy.**

		Task 1	Task 2
Predicate	Intrinsic	116	17
	Extrinsic	76	28
Entity	Intrinsic	128	28
	Extrinsic	115	62
Circumstance	Intrinsic	82	22
	Extrinsic	78	33
Coreference	-	98	1
Other	-	0	5
Total		693	196

Table 2: **BUMP dataset statistics.**

To post-process all data collected in Task 2, we manually assign an error type to each data point, based on our error type taxonomy in Task 1. Without informing annotators of any specific error types, we observe the rate that the "Other" label occurs is only 2.5% for Task 2 in Table 2. This confirms that the vast majority of errors produced by humans adhere to our proposed taxonomy. For more details on Task 2, please see Appendix B.

Remark. For both tasks, we ask annotators to introduce only *one* error (by editing the reference summary in a minimal way). We acknowledge that some reference summaries may be unfaithful in the first place; nevertheless, for both tasks, edited summaries are based on reference summaries, by which we ensure the edited summaries are always more unfaithful than reference summaries.

4 Meta-Evaluation of Faithfulness Evaluation Metrics

In this section, we first describe the faithfulness evaluation metrics benchmarked on BUMP (4.1).

Article (Partial)	Reference Summary	Required Error Type	Edited Summary	Corrected Error Type
... The drugs, whose value is estimated at more than **$105 million**, ... Officers arrested one Venezuelan and two Spanish citizens who were on board the vessel off the coast ... French customs officials seized nearly **250** kilograms (550 pounds) of cocaine on a vessel that was also off the coast of Martinique, according to authorities.	The value of the drugs is estimated at more than **$105** million. Officers arrested one Venezuelan and two Spanish citizens on board the vessel.	Intrinsic Circumstance Error	The value of the drugs is estimated at more than **$250** million . Officers arrested one Venezuelan and two Spanish citizens on board the vessel.	Intrinsic Entity Error
Lightning, floods and a deluge of hailstones descended on St Louis Tuesday as powerful storms **pummeled** the mid-United States. Roads around the Missouri city were flooded in the intense downpour, with one town recording more than two inches of rain in half an hour. ...	St Louis **was hit** Tuesday by flash floods. A nearby town had more than two inches of rain in less than half an hour.	Extrinsic Predicate Error	St Louis **pummeled** Tuesday by flash floods . A nearby town had more than two inches of rain in less than half an hour.	Intrinsic Predicate Error

Table 3: **Example error type corrections.** The above examples illustrate instances where annotators' edits to the reference do not match the *required error type* they are requested to produce. For such examples, BUMP includes a manually annotated *corrected error type*.

Then meta-evaluation protocols are discussed (4.2).

4.1 Faithfulness Metrics

To cover diverse types of faithfulness metrics, in this section, we select metrics that are generally used for measuring generation quality (i.e., n-gram-based metrics), recent metrics that are proposed specifically for faithfulness evaluations, as well as some pre-trained model based metrics, which are detailed as follows. We investigate their abilities to distinguish faithful summaries from their minimally edited counterparts.

• **n-Gram-Based Metrics:** We evaluate the following 2 n-gram-based metrics: BLEU (Papineni et al., 2002) and ROUGE (ROUGE-2 Precision specifically) (Lin, 2004).

• **Faithfulness Evaluation Metrics:** We evaluate the following 7 faithfulness evaluation metrics: QuestEval (Scialom et al., 2021), Q^2 (Honovich et al., 2021), QAFactEval (Fabbri et al., 2022), FactCC (Kryscinski et al., 2020), DAE (Goyal and Durrett, 2020), SummaC (Laban et al., 2022) and CoCo (Xie et al., 2021) in this paper. To obtain a score for FactCC, we take the classifier probability of the summary being faithful.

• **Other Metrics:** We evaluate the following 3 pre-trained model based metrics: BLEURT (Sellam et al., 2020) with the BLEURT-20 checkpoint (Pu et al., 2021), BERTScore (Zhang et al., 2020) (specifically the BERTScore-precision variant) using the DeBERTa-xl-MNLI model (He et al., 2021), and BARTScore (Yuan et al., 2021) with a BART (Lewis et al., 2020) model fine-tuned on the CNN/DailyMail dataset.

Note that for reference-based metrics, faithfulness scores are computed by treating the input article as the reference, and the reference/edited summary as the system output. We also normalize the direction of the metric score so that a higher score always corresponds to better faithfulness from the metric's view, e.g., FactCC predicts the probability that a summary is unfaithful, and so to obtain a faithfulness score, we take its complement.

4.2 Meta-Evaluation

Each faithfulness metric takes an article-summary pair and outputs a numerical faithfulness score. In our analysis, we measure faithfulness scores for both the reference summary as well as the human-annotated erroneous summary. We quantify the difference between faithfulness metrics on BUMP using two measurement protocols: **consistency** and **ROC AUC**. Originally introduced in Gabriel et al. (2021), consistency measures the success rate of a metric assigning a lower faithfulness score to the erroneous unfaithful summary. In contrast, ROC AUC instead measures the overall capability of a metric to discriminate faithful from unfaithful content for an input summary, and has previously been used by Honovich et al. (2022) for meta-evaluation. Although other metrics such as balanced accuracy have also been used to evaluate discriminability (Laban et al., 2022), we opt to use ROC AUC as it does not require determining a decision threshold.

5 Results

We report and analyze the performance of faithfulness metrics in this section using meta-evaluation protocol consistency and ROC AUC.

Consistency. The consistency studies of the two tasks[3] for all the metrics are reported in Table 4. In terms of the difficulty per error type, 1) for Task 1,

[3] Note that for Task 2, the error types with only a few samples (e.g., Coreference and Other) are *not* analyzed separately.

	BART	CoCo	DAE	QAFaEv	BERT	QuEv	BLEURT	SUMMAC	R-2	BLEU	Q^2	FACTCC
Overall	91.9	90.8	87.9	84.0	81.4	78.6	74.5	68.4	67.2	66.1	65.7	59.5
Intrinsic Predicate	96.6**	88.8	87.1	79.3	81.0	69.0	69.8	61.2	51.7	39.7	49.1	49.1
Extrinsic Predicate	94.7	93.4	84.2	86.8	88.2	73.7	73.7	67.1	63.2	61.8	59.2	60.5
Intrinsic Entity	93.0	91.4	88.3	91.4	80.5	82.8	76.6	75.8	65.6	60.9	75.0	61.7
Extrinsic Entity	97.4	96.5	93.9	90.4	88.7	87.0	90.4	79.1	86.1	87.0	80.9	59.1
Intrinsic Circumstance	85.4	84.2	86.6	81.7	67.1	74.4	63.4	74.4	51.2	53.7	68.3	63.4
Extrinsic Circumstance	85.9	85.9	85.9	85.9	79.5	78.2	75.6	73.1	79.5	74.4	68.0	59.0
Coreference Error	86.7	92.9	86.7	70.4	82.7	82.7	67.4	46.9	72.5	86.7	56.1	65.3
Intrinsic Error	92.3*	88.7	87.4	84.7	77.3	75.8	70.9	70.3	57.1	51.5	64.1	57.7
Extrinsic Error	93.3	92.6	88.9	88.1	85.9	80.7	81.4	74.0	77.7	76.2	71.0	59.5

(a) Task 1

	BART	QAFaEv	CoCo	BERT	BLEURT	DAE	QuEv	SUMMAC	R-2	BLEU	Q^2	FACTCC
Overall	93.4**	85.7	84.7	82.1	77.6	75.5	75.5	73.0	68.9	66.8	65.8	48.0
Intrinsic Predicate	82.4	88.2	88.2	82.4	82.4	82.4	82.4	64.7	70.6	64.7	76.5	64.7
Extrinsic Predicate	92.9	92.9	89.3	85.7	75.0	64.3	67.9	89.3	71.4	53.6	57.1	42.9
Intrinsic Entity	96.4*	78.6	78.6	82.1	67.9	64.3	60.7	78.6	53.6	50.0	42.9	39.3
Extrinsic Entity	95.2	88.7	85.5	80.7	80.7	79.0	80.7	79.0	69.4	72.6	67.7	45.2
Intrinsic Circumstance	90.9	81.8	81.8	72.7	72.7	63.6	77.3	59.1	59.1	77.3	68.2	63.6
Extrinsic Circumstance	97.0	78.8	87.9	87.9	81.8	93.9	75.8	54.6	81.8	75.8	81.8	48.5
Intrinsic Error	91.0	82.1	82.1	79.1	73.1	68.7	71.6	68.7	59.7	62.7	59.7	53.7
Extrinsic Error	95.1**	87.0	87.0	83.7	79.7	79.7	76.4	74.8	73.2	69.1	69.1	45.5

(b) Task 2

Table 4: **Consistency (%) of faithfulness evaluation metrics.** BART: BARTSCORE, QAFaEv: QAFACTEVAL, BERT: BERTSCORE, QuEv: QUESTEVAL, R-2: ROUGE-2. All values are color-coded. For each row, $*$ ($p < 0.05$) and $**$ ($p < 0.01$) indicate the results are statistically significant when comparing the best to the second-best metric.

*Extrinsic Entity Error*s are generally the easiest, while all but BARTSCORE struggle with *Intrinsic Predicate Error*s; 2) for Task 2, *Intrinsic Entity Error*s are the hardest. This implies that when annotators are not presented with any error types, the introduced error styles may differ from those in Task 1 (see Section 6), potentially causing inconsistencies for metrics in these two tasks. Nevertheless, we observe that for both tasks, *Intrinsic Error*s are more challenging than extrinsic ones across all but FACTCC in Task 2. This is likely because *Intrinsic Error*s can be derived from the original article, while *Extrinsic Error*s contain words that do not appear in the original article, making *Intrinsic Error*s more subtle to be identified than extrinsic ones.

For the overall performance (all error types are considered), BARTSCORE has the highest consistency in both tasks, though BARTSCORE has not been proposed specifically for faithfulness evaluations. Other metrics that rank top 4 in both tasks include QAFACTEVAL and COCO. By comparison, Q^2 and FACTCC have the worst consistency, even worse than n-gram-based metrics ROUGE and BLEU; nevertheless, they exhibit different rankings in terms of ROC AUC (see the next section).

ROC AUC. ROC AUC scores are presented in Table 5. We observe that the overall ranking of faithfulness metrics according to ROC AUC substantially differs from the ranking according to consistency. In particular, the rank of BARTSCORE drops from the top one to the fifth, while Q^2 improves significantly from second to last to second overall. QAFACTEVAL consistently exhibits high performance and even ranks first under ROC AUC, while n-gram based metrics, e.g., ROUGE-2 and BLEU consistently show the worst performance, as expected. In general, metrics that are specifically proposed for faithfulness evaluations rank higher than generic NLG evaluation metrics.

We additionally observe that the relative rankings of ROC AUC scores across error types and task settings are largely consistent with the relative rankings of consistency scores. Specifically, we again observe that on a per metric basis: 1) ROC AUC scores are generally lower for Task 2 than Task 1 (particularly for *Entity Error*s), and 2) metrics generally show worse performance on *Intrinsic Error*s than extrinsic ones.

For our two meta-evaluation protocols, consistency is suitable for the pairwise ranking of two summaries for a given input article, while ROC AUC is more adequate in evaluating the absolute capacity of unfaithful summary detection. If a metric has high consistency but low ROC AUC, it implies that the scores for predicted faithful and unfaithful summaries overlap frequently. Such overlap makes

	QAFaEv	Q^2	DAE	QuEv	BART	FactCC	CoCo	SummaC	BLEURT	BERT	R-2	BLEU
Overall	71.5**	64.2	63.7	62.0	60.1	57.2	56.4	55.9	55.1	55.0	53.2	50.6
Intrinsic Predicate	66.7**	57.0	60.4	56.0	60.7	50.1	55.0	53.4	54.3	55.2	51.6	50.3
Extrinsic Predicate	73.5**	59.6	58.7	59.8	58.1	50.8	55.5	57.4	54.2	56.8	53.3	50.5
Intrinsic Entity	77.7*	71.7	70.1	67.8	63.9	62.9	58.4	58.0	56.0	56.3	53.1	50.6
Extrinsic Entity	78.4*	73.1	67.7	71.2	63.5	57.0	58.6	57.5	58.7	56.8	55.4	50.8
Intrinsic Circumstance	71.3*	62.9	63.7	55.3	55.4	61.9	53.8	57.2	52.2	51.0	51.9	49.8
Extrinsic Circumstance	73.9**	64.0	61.0	57.5	57.0	58.8	54.8	59.3	53.4	52.4	55.1	50.6
Coreference Error	58.2	57.0	60.8	62.1	59.7	58.0	57.2	50.1	55.4	55.3	53.0	51.7
Intrinsic Error	72.1**	64.4	64.9	60.4	60.5	58.1	56.0	56.1	54.4	54.6	52.3	50.3
Extrinsic Error	75.6**	66.5	63.2	63.9	59.8	55.7	56.5	57.9	55.8	55.4	54.6	50.6

(a) Task 1

	QAFaEv	Q^2	DAE	QuEv	BART	SummaC	CoCo	BERT	R-2	BLEURT	FactCC	BLEU
Overall	71.2**	61.3	58.8	57.4	57.4	56.9	54.5	54.1	54.0	52.6	51.5	50.3
Intrinsic Predicate	72.3	63.3	60.6	57.4	55.0	55.9	55.7	57.4	53.6	54.7	58.8	51.6
Extrinsic Predicate	74.7**	57.1	57.5	55.6	60.2	59.8	55.9	56.0	55.2	54.3	48.3	50.3
Intrinsic Entity	65.3	56.6	57.8	55.5	60.2	57.7	57.7	56.1	52.5	52.6	50.3	50.4
Extrinsic Entity	74.5**	64.6	58.1	60.4	58.7	59.6	55.3	54.1	54.8	53.0	49.7	50.7
Intrinsic Circumstance	65.2	57.0	55.4	55.8	56.6	55.9	53.9	51.9	52.4	51.9	51.4	51.3
Extrinsic Circumstance	71.9	66.0	67.9	57.3	57.9	54.1	55.0	53.8	55.4	54.5	55.6	51.2
Intrinsic Error	66.5**	57.6	57.5	55.4	55.7	56.0	54.6	53.8	52.2	51.5	52.7	50.3
Extrinsic Error	73.5**	63.4	59.8	58.4	58.5	57.9	54.8	54.1	54.7	53.4	50.9	50.4

(b) Task 2

Table 5: **ROC AUC (%) of faithfulness evaluation metrics.** BART: BARTScore, QAFaEv: QAFactEval, BERT: BERTScore, QuEv: QuestEval, R-2: ROUGE-2. All values are color-coded. For each row, $*$ ($p < 0.05$) and $**$ ($p < 0.01$) indicate the results are statistically significant when comparing the best to the second-best metric.

it challenging to establish a clear decision boundary for classifications. Hence, to improve the classification capability of metrics with high consistency, more calibration is needed to increase the score gap between faithful and unfaithful summaries.

6 Analysis of BUMP

In this section, we conduct more analysis of BUMP by studying how BUMP differs from other benchmarks, followed by a qualitative analysis of the detection difficulty between Tasks 1 and 2.

Comparison with Model-Generated Unfaithful Summaries. We compare the generation probabilities of our edited summaries to those of summaries generated from beam search by a BART-based summarizer (trained using the training data of CNN/DailyMail) for the same set of documents in our dataset. We report the difference of these generation probabilities normalized by the text length in Figure 2, where we find our edited summaries are much different from model generations in terms of the model generation probabilities. This highlights that existing metrics may not work well on summaries of various styles and experiments are needed to verify their effectiveness in human-generated unfaithful summaries.

Furthermore, we compare our ROC AUC scores with those in existing datasets as shown in TRUE

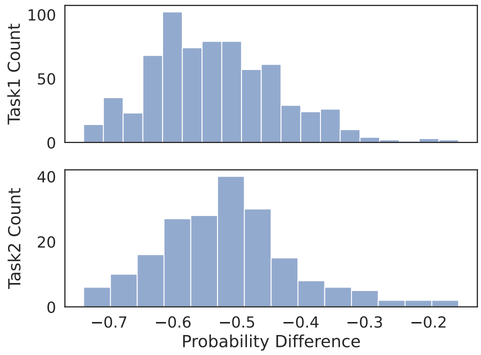


Figure 2: **Distribution of probability differences between human-edited and model-generated summaries on Tasks 1 and 2**. Probabilities are computed using a BART-based summarizer. The high frequency of negative values indicates that human-edited summaries tend to have lower probabilities under the model.

(Honovich et al., 2022). In BUMP, faithful and unfaithful samples under each error type are balanced for both Tasks 1 and 2. Therefore, for a fair comparison, we pick QAGS-C (Wang et al., 2020) (also a balanced dataset on CNN/DailyMail) in TRUE. In Table 5, it shows that the ROC AUC scores evaluated on BUMP are generally much smaller (50–70% with many values close to random baseline), whereas most ROC AUC scores are 70–84% in QAGS-C (see Appendix C). This again

Article (Partial)	Reference Summary	Error Type	Edited Summary	
			Task 1	Task 2
... Detective Chief Inspector Paul Johnson of the London Metropolitan Police Flying Squad said the thieves appeared to have gained access to the vault of Hatton Garden Safe Deposit Ltd through the shaft of **an elevator that is used by several businesses in the building**. ...	Police say the thieves gained entry through the building's **communal** elevator shaft. ...	Extrinsic Entity Error	Police say the thieves gained entry through the building's communal **staircase**. ...	Police say the thieves gained entry through the building's **private** elevator shaft. ...
... **Claire Nugent**, 43, and Nigel Morter, 47, have been married for 14 years... She said: 'Every night I come home to my Sixties bubble, switch on my old **record player**, listen to some vinyl, and all the stresses of 2015 melt away' ...	**Claire Nugent** and Nigel Morter restored a ... likes to come home and switch on an **old record player** like in the 60s.	Extrinsic Entity Error	**Tim Horton** and Nigel Morter restored a ... likes to come home and switch on an old record player like in the 60s.	Claire Nugent and Nigel Morter restored a ... likes to come home and switch on a **black and white TV** like in the 60s.
Almost three years after nearly leaving Liverpool Jordan Henderson **has committed** his long-term future to the club convinced he can win silverware at Anfield. ... Henderson has **urged** Liverpool team-mate Raheem Sterling to follow his lead by signing a new deal.	Jordan Henderson **has signed** a new five-year deal at Anfield. ... Henderson has **urged** Raheem Sterling to ...	Extrinsic Predicate Error	Jordan Henderson has signed a new five-year deal at Anfield. ... Henderson has **discouraged** Raheem Sterling to ...	Jordan Henderson **is considering signing** a new five-year deal at Anfield. ... Henderson has urged Raheem Sterling to ...

Table 6: **Qualitative examples illustrating the higher difficulty of edits in Task 2.** Each row contains a pair of edits from Tasks 1 and 2 pertaining to the same article and error type, where more metrics are inconsistent for Task 2 edits.

indicates that the human-generated errors in BUMP are more difficult for metrics to detect than model-generated errors in existing datasets, reinforcing the value of BUMP as a challenging benchmark for evaluating faithfulness metrics. In addition, we also compare the ROC AUC rankings of different faithfulness metrics under QAGS-C and BUMP. Specifically, we summarize the performance rankings under QAGS-C from Appendix C as well as those from Table 5 under Intrinsic/Extrinsic Error types in Tasks 1 and 2, and report them in Figure 3, where only faithfulness metrics used in both (Honovich et al., 2022) and Table 5 are presented. In Figure 3, we observe that for some faithfulness metrics, such as Q^2 and BARTSCORE, their ROC AUC rankings are quite stable across all datasets. However, for other faithfulness metrics, the performance ranking under QAGS-C is very different from the ranking derived from BUMP, e.g., QUESTEVAL mostly exhibits high ROC AUC ranking in BUMP; by contrast, it experiences the worst performance in QAGS-C. Thus, we believe BUMP complements existing benchmarks and allows a more comprehensive analysis of faithfulness metrics in future studies.

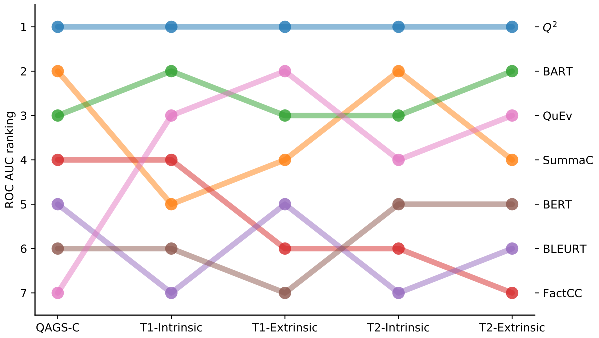


Figure 3: **ROC AUC ranking of faithfulness metrics under different datasets**. Only faithfulness metrics used in both (Honovich et al., 2022) and Table 5 are presented. T1: Task 1 of BUMP, T2: Task 2 of BUMP, Intrinsic: Intrinsic Errors, Extrinsic: Extrinsic Errors, BART: BARTSCORE, BERT: BERTSCORE, QuEv: QUESTEVAL.

Qualitative Analysis. We provide a qualitative analysis of examples that demonstrate the increased difficulty of Task 2. The examples are provided in Table 6. Each row contains edited summaries from Tasks 1 and 2 for the same original article and its reference summary. In addition, to compare edited summaries under the same error type, we pick examples where the corrected error type from Task 1 is the same as the exhibited error type from Task 2. As shown in Table 6, in the first example, for the *Extrinsic Entity Error* type, the annotator in Task 1 modifies the entity *elevator shaft* to another entity *staircase*. Whereas the annotator in Task 2 modifies the word *communal* to *private* (i.e., also an *Extrinsic Entity Error*) which requires commonsense knowledge to infer that *private* is contradictory to the fact that *the elevator is used by several businesses in the building*. In the second example, for the *Extrinsic Entity Error* type, the annotator in Task 1 modifies the entity name from *Claire Nugent* to a random name *Tim Horton*, whereas the annotator in Task 2 changes *record player* to *black and white TV* to fit the *60s* theme, which again, requires additional knowledge. In the last example, the annotator in Task 2 modifies the temporal state of the action *sign* from *signed* to *is considering signing* which is more challenging than changing the action *urged* to its antonym *discouraged* as the annotator in Task 1 does.

For the first two examples of Task 2 in Table 6, only 4 metrics (QAFactEval, QuestEval, BLEURT, and BERTScore for the first example; QAFactEval, SummaC, BLEURT, and ROUGE-2 for the second example) succeed in giving a higher score to the reference summary. In comparison, 9 and 11 metrics succeed in giving a higher score to the reference summary in their Task 1 counterparts, respectively. For the last example, 8 metrics succeed in Task 2 and all 12 metrics succeed in Task 1. Thus, Table 6 shows that some unfaithful summaries in Task 2 are more challenging for faithfulness metrics to detect, which further exemplifies the challenges of Task 2 in BUMP.

7 Conclusion

In this paper, we presented a benchmark of unfaithful minimal pairs (BUMP) to evaluate faithfulness metrics. Unlike prior work where all unfaithful summaries are model generated, each unfaithful summary in BUMP is generated by minimal human edits to introduce one unfaithful error given a reference summary. Through our experiments, we found that BUMP complements existing benchmarks in a number of ways. First, we found that the summaries in BUMP are harder to discriminate and less probable under SOTA summarization models. Second, we used BUMP to measure the consistency of metrics, which cannot be readily measured using other benchmarks. This analysis revealed a discrepancy between the discriminability and consistency of existing metrics, highlighting an important area for future faithfulness metric research to address. Finally, we used BUMP to study faithfulness metrics' performance on individual error types—where our minimal-pair-inspired setup helped control for conclusions being conflated across multiple error types—which revealed that sensitivity to intrinsic errors is another important area for future research to focus on.

Acknowledgements

We would like to thank our colleague Aoife Cahill at Dataminr for her valuable comments, suggestions, and support for this paper. We also thank the anonymous reviewers for their feedback and comments.

Limitations

Although BUMP is, to our knowledge, the first dataset on which to study the consistency of faithfulness metrics on human-written errors across different error types, there are some limitations regarding the conclusions that can be drawn from it. For one, because BUMP is comprised of minimal edits to reference summaries from CNN/DailyMail, it is not suitable for analyzing the consistency of faithfulness metrics when errors are added to reference summaries already containing many errors. In addition, due to a combination of resource constraints and human preferences for writing specific types of errors, the sample sizes for some error types in Task 2 (e.g., *Coreference Error* and *Intrinsic Predicate Error*) may not be sufficiently large to enable statistically significant comparisons between different metrics for specific error types.

Ethics Statement

The collection of BUMP involves human annotations. The human annotators are provided with clear task instructions and informed of the conditions where they would be qualified and disqualified. We compensate annotators with \$3.00 per assignment in the qualification task and \$0.50 per assignment in the full task for both Tasks 1 and 2. The final paid rate is \$15 per hour which is over the US national minimum wage[4] of \$7.25. We are also aware that our shared datasets could be potentially misused as training samples, albeit a small number, to develop models to generate unfaithful content.

References

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo-*

[4] https://www.dol.gov/general/topic/wages/minimumwage

gies, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Saadia Gabriel, Asli Celikyilmaz, Rahul Jha, Yejin Choi, and Jianfeng Gao. 2021. GO FIGURE: A meta evaluation of factuality in summarization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 478–487, Online. Association for Computational Linguistics.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DEBERTA: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 161–175, Dublin, Ireland. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alexander R. Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F. Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

A Details of Task 1: Taxonomy-based Unfaithful Summaries

A.1 Qualification Task

The instructions and the task interface for the qualification task of Task 1 are shown in Figures A1 to A4.

In this qualification task, all US-based annotators are able to participate. Specifically, we ask annotators to read a news article and seven pairs of summaries. For each pair of summaries, the first summary is the correct reference summary, and the second summary is the unfaithfully edited summary that contains one of the seven error types in our taxonomy. We then ask the annotators to select one answer from the seven error types to indicate which type of error is introduced in the edited unfaithful summary. Only the annotators who answered 6 out of these 7 questions correctly passed the qualification task. We launched 3 batches in total with 9 assignments for each batch, and 9 annotators passed the qualification task.

A.2 Full Task

The instructions and the task interface for the full task of Task 1 are shown in Figures A5 to A7.

In the full task for Task 1, different from the qualification task, we ask the annotators to read a news article from CNN/DailyMail (Hermann et al., 2015) and one reference summary for the article. We then ask the annotators to edit the reference summary to introduce the error type specified through a minimal edit. If they cannot introduce the error type based on the reference summary, they can write "N/A" to indicate that it is impossible to introduce the specified error type based on the provided reference summary. There are 18 samples in Task 1 dataset that are annotated as "N/A" by the annotators, all of which are reviewed by the authors of this paper and re-annotated with the correct edits (we note that the required error types can be provided for all these cases) as a post-processing step to ensure the completeness of the dataset.

In addition, for Task 1, to help reduce the confusion from annotators regarding *Circumstance Error*s and *Entity Error*s, we explicitly specify that the *Circumstance Error*s should only be erroneous edits concerning the time, duration, or location of an event, and changing the quantity of a *noun* is not considered as a *Circumstance Error*.

B Details of Task 2: Freestyle Unfaithful Summaries

B.1 Qualification Task

The instructions and the task interface for the qualification task of Task 2 are shown in Figures A8 to A9.

In this qualification task, all US-based annotators who did not participate in the qualification task of Task 1 are qualified to participate. Specifically, we show the annotators four pairs of news article and its summary from CNN/DailyMail, and ask them to answer if the summaries are faithful based on the original news articles. Among the four pairs, three of them are unfaithful and one is faithful. Only the annotators who answered correctly to all of these 4 pairs passed the qualification task. We launched 3 batches in total with 9 assignments for each batch, and 8 annotators passed the qualification task.

B.2 Full Task

The instructions and the task interface for the full task of Task 2 are shown in Figures A10 to A11.

In the full task of Task 2, unlike Task 1, we do not list any potential error types so as to achieve freestyle editing. The edited summary is valid as long as only one error is introduced based on the reference summary via a minimal edit. Furthermore, we also do the following to ensure the quality of edited summaries:

- For minimal edits, we explicitly ask annotators *not* to write from scratch, but to introduce only one error on top of the given reference summary.

- In the pilot study, we notice that some edited summaries are simply removing/adding sentences or phrases (such data points are removed in the final released data); we, therefore, add additional instructions that require the edited and the reference summaries to contain a similar amount of information about the given news article (i.e., similar coverage).

- The edited summaries should be grammatically correct.

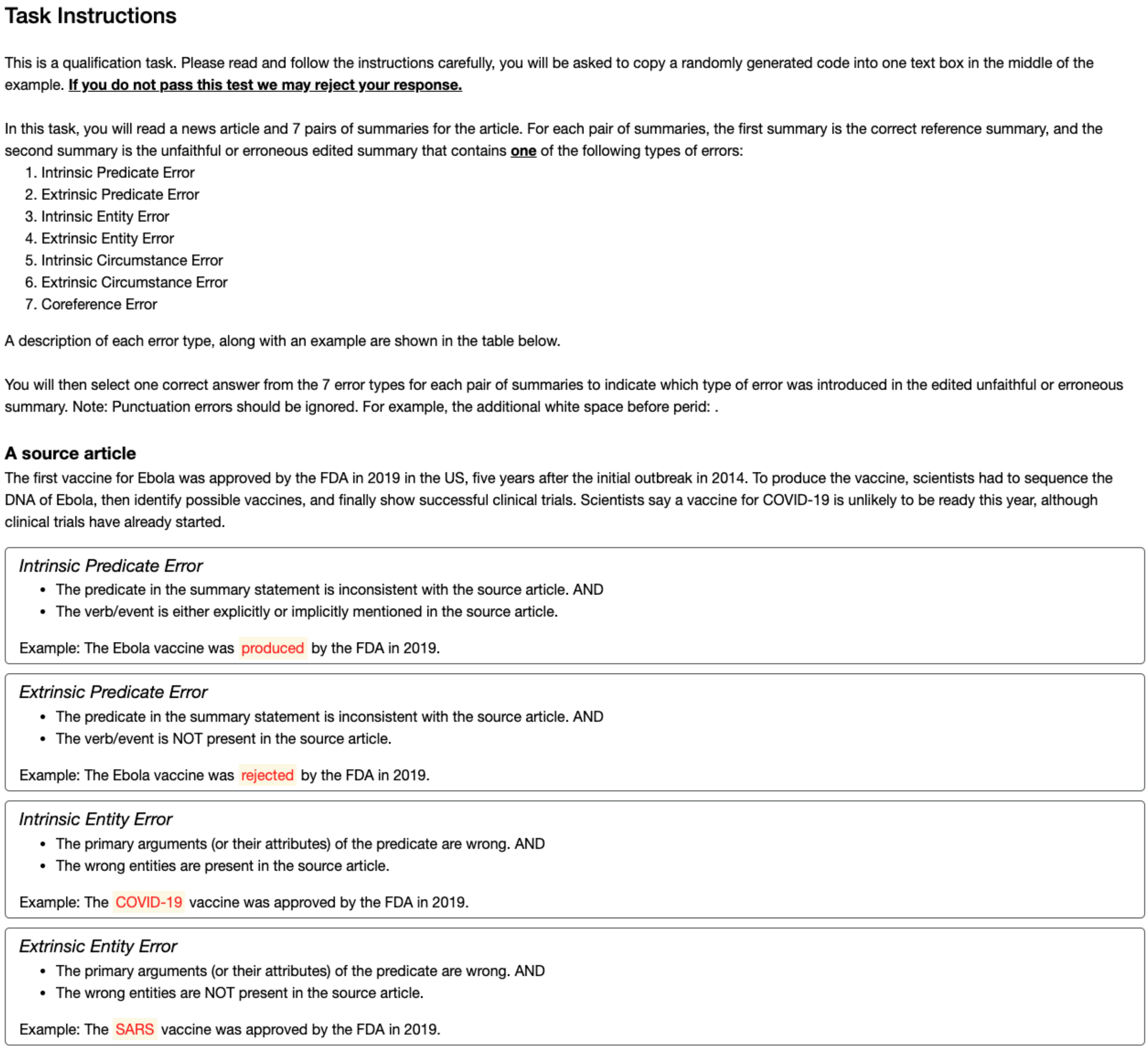
- The edited summaries should be plausible and adhere to common sense.

- Some examples of edited summaries are given in the task instructions.

C ROC AUC Results from Other Benchmarks

To compare BUMP with other benchmarks, we also report the ROC AUC scores from TRUE (Honovich et al., 2022). Specifically, in BUMP, faithful and unfaithful samples under each error type are balanced for both Tasks 1 and 2. Therefore, for a fair comparison, 1) in TRUE, we pick QAGS-C (Wang et al., 2020), which is also a balanced dataset on CNN/DailyMail. The ROC AUC scores of QAGS-C are reported in Table A1; 2) for faithfulness metrics in Table A1, we use the same implementation and model checkpoints in this paper as those in TRUE (Honovich et al., 2022). Then according to Table A1, the metric performance ranking in terms of ROC AUC for QAGS-C is Q^2 > SummaC > BARTScore > FactCC > BLEURT > BERTScore > QuestEval, which is very different from the ranking derived from our BUMP dataset, e.g., SummaC exhibits worse ROC AUC than QuestEval for most error types in both Tasks 1 and 2 (see Table 5).

In addition to the balanced dataset QAGS-C in TRUE, we also report the ROC AUC scores of imbalanced FRANK (Pagnoni et al., 2021) and SummEval (Fabbri et al., 2021) datasets (two datasets containing CNN/DailyMail) from TRUE in Table A1. Although the FRANK and SummEval datasets are imbalanced, we have similar observations as those from the QAGS-C dataset: 1) their ROC AUC scores (mostly 70–90%) are much larger than the ROC AUC scores (50–70%) derived from our BUMP dataset; 2) in terms of the ROC AUC ranking, the top two remain Q^2 and SummaC for both FRANK and SummEval, and SummaC always ranks higher than QuestEval. By contrast, in Table 5, we show that SummaC mostly exhibits worse ROC AUC than QuestEval.



Task Instructions

This is a qualification task. Please read and follow the instructions carefully, you will be asked to copy a randomly generated code into one text box in the middle of the example. **If you do not pass this test we may reject your response.**

In this task, you will read a news article and 7 pairs of summaries for the article. For each pair of summaries, the first summary is the correct reference summary, and the second summary is the unfaithful or erroneous edited summary that contains **one** of the following types of errors:

1. Intrinsic Predicate Error
2. Extrinsic Predicate Error
3. Intrinsic Entity Error
4. Extrinsic Entity Error
5. Intrinsic Circumstance Error
6. Extrinsic Circumstance Error
7. Coreference Error

A description of each error type, along with an example are shown in the table below.

You will then select one correct answer from the 7 error types for each pair of summaries to indicate which type of error was introduced in the edited unfaithful or erroneous summary. Note: Punctuation errors should be ignored. For example, the additional white space before perid: .

A source article

The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.

Intrinsic Predicate Error

- The predicate in the summary statement is inconsistent with the source article. AND
- The verb/event is either explicitly or implicitly mentioned in the source article.

Example: The Ebola vaccine was produced by the FDA in 2019.

Extrinsic Predicate Error

- The predicate in the summary statement is inconsistent with the source article. AND
- The verb/event is NOT present in the source article.

Example: The Ebola vaccine was rejected by the FDA in 2019.

Intrinsic Entity Error

- The primary arguments (or their attributes) of the predicate are wrong. AND
- The wrong entities are present in the source article.

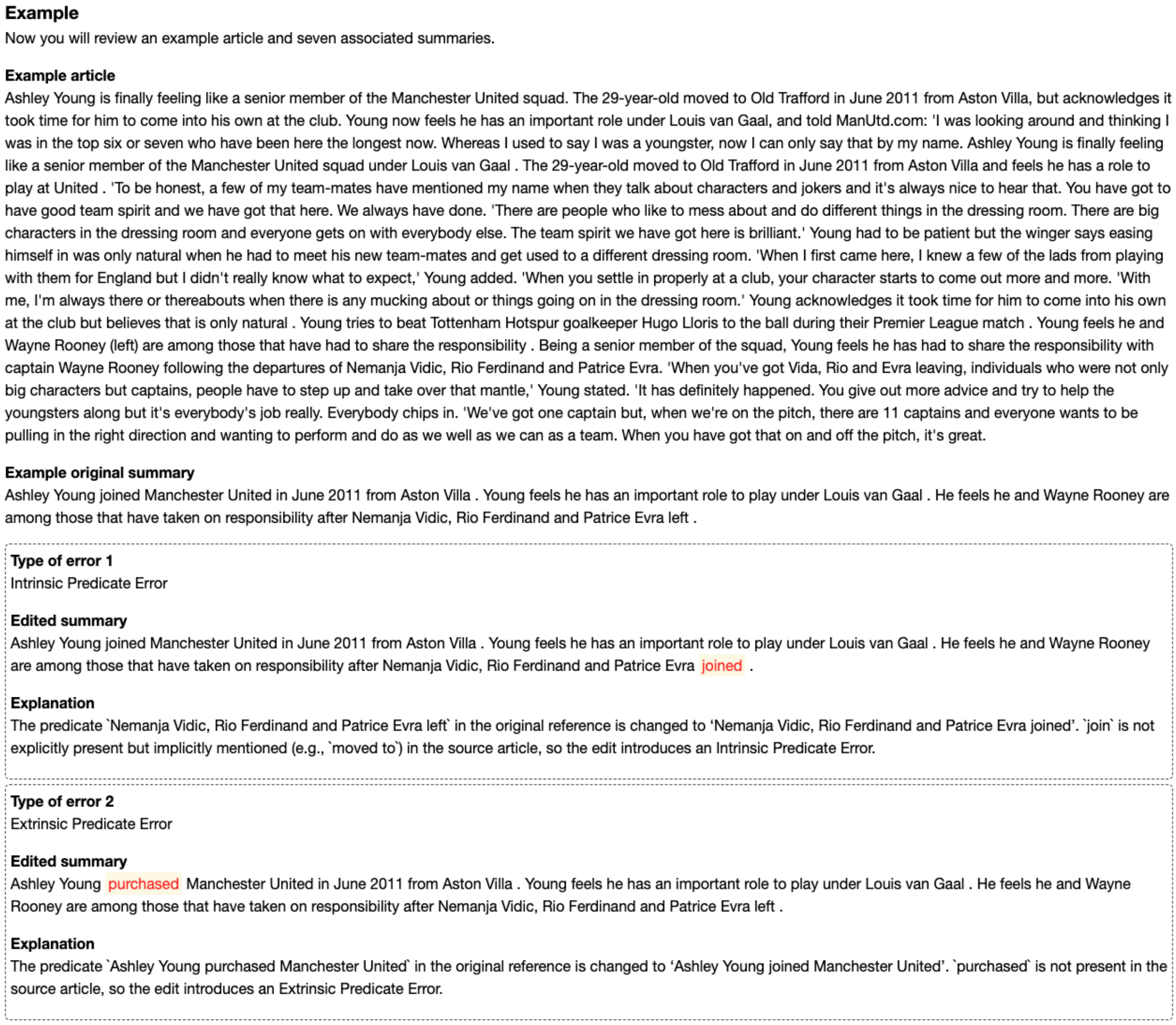
Example: The COVID-19 vaccine was approved by the FDA in 2019.

Extrinsic Entity Error

- The primary arguments (or their attributes) of the predicate are wrong. AND
- The wrong entities are NOT present in the source article.

Example: The SARS vaccine was approved by the FDA in 2019.

Figure A1: Screenshot of the qualification task for Task 1 (1/4).



Example

Now you will review an example article and seven associated summaries.

Example article

Ashley Young is finally feeling like a senior member of the Manchester United squad. The 29-year-old moved to Old Trafford in June 2011 from Aston Villa, but acknowledges it took time for him to come into his own at the club. Young now feels he has an important role under Louis van Gaal, and told ManUtd.com: 'I was looking around and thinking I was in the top six or seven who have been here the longest now. Whereas I used to say I was a youngster, now I can only say that by my name. Ashley Young is finally feeling like a senior member of the Manchester United squad under Louis van Gaal . The 29-year-old moved to Old Trafford in June 2011 from Aston Villa and feels he has a role to play at United . 'To be honest, a few of my team-mates have mentioned my name when they talk about characters and jokers and it's always nice to hear that. You have got to have good team spirit and we have got that here. We always have done. 'There are people who like to mess about and do different things in the dressing room. There are big characters in the dressing room and everyone gets on with everybody else. The team spirit we have got here is brilliant.' Young had to be patient but the winger says easing himself in was only natural when he had to meet his new team-mates and get used to a different dressing room. 'When I first came here, I knew a few of the lads from playing with them for England but I didn't really know what to expect,' Young added. 'When you settle in properly at a club, your character starts to come out more and more. 'With me, I'm always there or thereabouts when there is any mucking about or things going on in the dressing room.' Young acknowledges it took time for him to come into his own at the club but believes that is only natural . Young tries to beat Tottenham Hotspur goalkeeper Hugo Lloris to the ball during their Premier League match . Young feels he and Wayne Rooney (left) are among those that have had to share the responsibility . Being a senior member of the squad, Young feels he has had to share the responsibility with captain Wayne Rooney following the departures of Nemanja Vidic, Rio Ferdinand and Patrice Evra. 'When you've got Vida, Rio and Evra leaving, individuals who were not only big characters but captains, people have to step up and take over that mantle,' Young stated. 'It has definitely happened. You give out more advice and try to help the youngsters along but it's everybody's job really. Everybody chips in. 'We've got one captain but, when we're on the pitch, there are 11 captains and everyone wants to be pulling in the right direction and wanting to perform and do as we well as we can as a team. When you have got that on and off the pitch, it's great.

Example original summary

Ashley Young joined Manchester United in June 2011 from Aston Villa . Young feels he has an important role to play under Louis van Gaal . He feels he and Wayne Rooney are among those that have taken on responsibility after Nemanja Vidic, Rio Ferdinand and Patrice Evra left .

Type of error 1

Intrinsic Predicate Error

Edited summary

Ashley Young joined Manchester United in June 2011 from Aston Villa . Young feels he has an important role to play under Louis van Gaal . He feels he and Wayne Rooney are among those that have taken on responsibility after Nemanja Vidic, Rio Ferdinand and Patrice Evra joined .

Explanation

The predicate `Nemanja Vidic, Rio Ferdinand and Patrice Evra left` in the original reference is changed to 'Nemanja Vidic, Rio Ferdinand and Patrice Evra joined'. `join` is not explicitly present but implicitly mentioned (e.g., `moved to`) in the source article, so the edit introduces an Intrinsic Predicate Error.

Type of error 2

Extrinsic Predicate Error

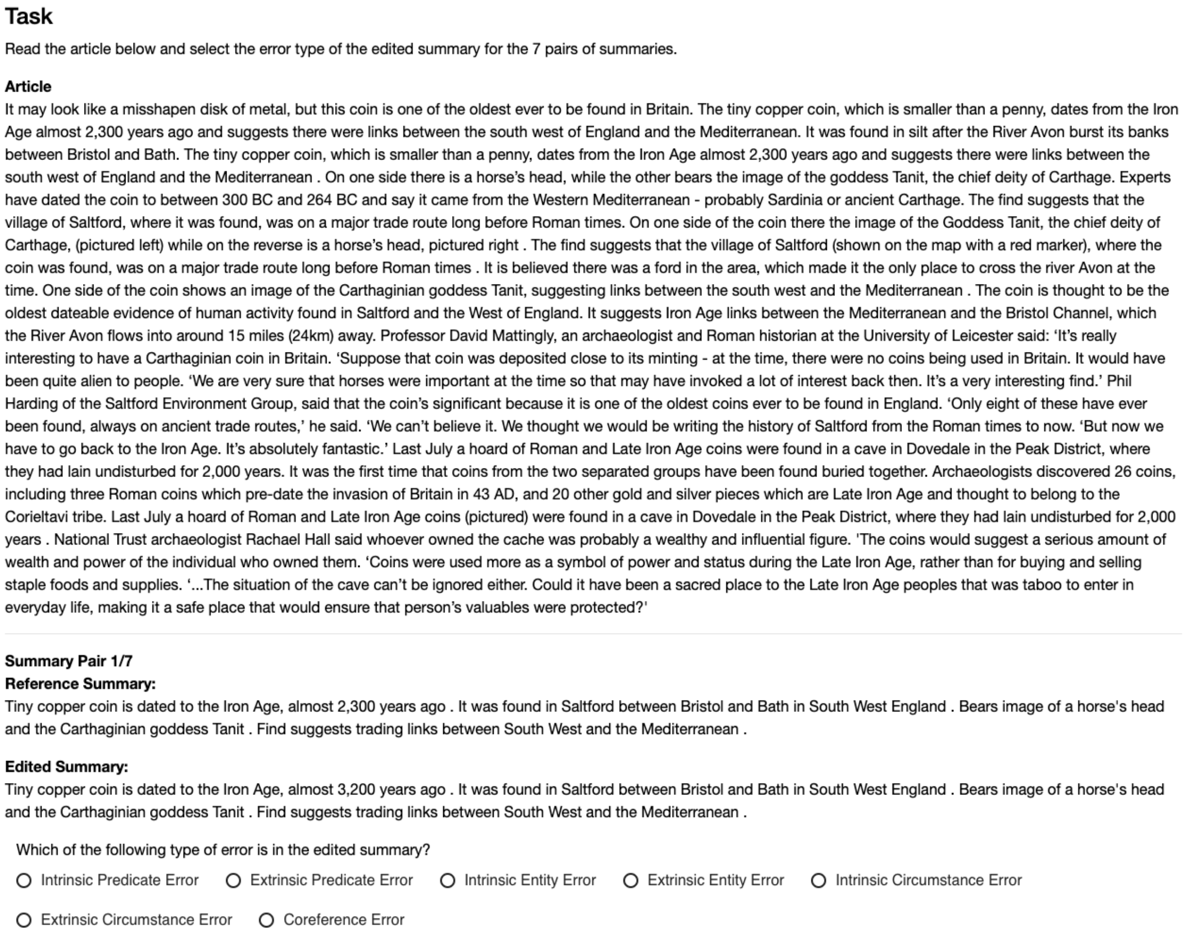
Edited summary

Ashley Young purchased Manchester United in June 2011 from Aston Villa . Young feels he has an important role to play under Louis van Gaal . He feels he and Wayne Rooney are among those that have taken on responsibility after Nemanja Vidic, Rio Ferdinand and Patrice Evra left .

Explanation

The predicate `Ashley Young purchased Manchester United` in the original reference is changed to 'Ashley Young joined Manchester United'. `purchased` is not present in the source article, so the edit introduces an Extrinsic Predicate Error.

Figure A2: Screenshot of the qualification task for Task 1 (2/4).



Task

Read the article below and select the error type of the edited summary for the 7 pairs of summaries.

Article

It may look like a misshapen disk of metal, but this coin is one of the oldest ever to be found in Britain. The tiny copper coin, which is smaller than a penny, dates from the Iron Age almost 2,300 years ago and suggests there were links between the south west of England and the Mediterranean. It was found in silt after the River Avon burst its banks between Bristol and Bath. The tiny copper coin, which is smaller than a penny, dates from the Iron Age almost 2,300 years ago and suggests there were links between the south west of England and the Mediterranean . On one side there is a horse's head, while the other bears the image of the goddess Tanit, the chief deity of Carthage. Experts have dated the coin to between 300 BC and 264 BC and say it came from the Western Mediterranean - probably Sardinia or ancient Carthage. The find suggests that the village of Saltford, where it was found, was on a major trade route long before Roman times. On one side of the coin there the image of the Goddess Tanit, the chief deity of Carthage, (pictured left) while on the reverse is a horse's head, pictured right . The find suggests that the village of Saltford (shown on the map with a red marker), where the coin was found, was on a major trade route long before Roman times . It is believed there was a ford in the area, which made it the only place to cross the river Avon at the time. One side of the coin shows an image of the Carthaginian goddess Tanit, suggesting links between the south west and the Mediterranean . The coin is thought to be the oldest dateable evidence of human activity found in Saltford and the West of England. It suggests Iron Age links between the Mediterranean and the Bristol Channel, which the River Avon flows into around 15 miles (24km) away. Professor David Mattingly, an archaeologist and Roman historian at the University of Leicester said: 'It's really interesting to have a Carthaginian coin in Britain. 'Suppose that coin was deposited close to its minting - at the time, there were no coins being used in Britain. It would have been quite alien to people. 'We are very sure that horses were important at the time so that may have invoked a lot of interest back then. It's a very interesting find.' Phil Harding of the Saltford Environment Group, said that the coin's significant because it is one of the oldest coins ever to be found in England. 'Only eight of these have ever been found, always on ancient trade routes,' he said. 'We can't believe it. We thought we would be writing the history of Saltford from the Roman times to now. 'But now we have to go back to the Iron Age. It's absolutely fantastic.' Last July a hoard of Roman and Late Iron Age coins were found in a cave in Dovedale in the Peak District, where they had lain undisturbed for 2,000 years. It was the first time that coins from the two separated groups have been found buried together. Archaeologists discovered 26 coins, including three Roman coins which pre-date the invasion of Britain in 43 AD, and 20 other gold and silver pieces which are Late Iron Age and thought to belong to the Corieltavi tribe. Last July a hoard of Roman and Late Iron Age coins (pictured) were found in a cave in Dovedale in the Peak District, where they had lain undisturbed for 2,000 years . National Trust archaeologist Rachael Hall said whoever owned the cache was probably a wealthy and influential figure. 'The coins would suggest a serious amount of wealth and power of the individual who owned them. 'Coins were used more as a symbol of power and status during the Late Iron Age, rather than for buying and selling staple foods and supplies. '...The situation of the cave can't be ignored either. Could it have been a sacred place to the Late Iron Age peoples that was taboo to enter in everyday life, making it a safe place that would ensure that person's valuables were protected?'

Summary Pair 1/7
Reference Summary:
Tiny copper coin is dated to the Iron Age, almost 2,300 years ago . It was found in Saltford between Bristol and Bath in South West England . Bears image of a horse's head and the Carthaginian goddess Tanit . Find suggests trading links between South West and the Mediterranean .

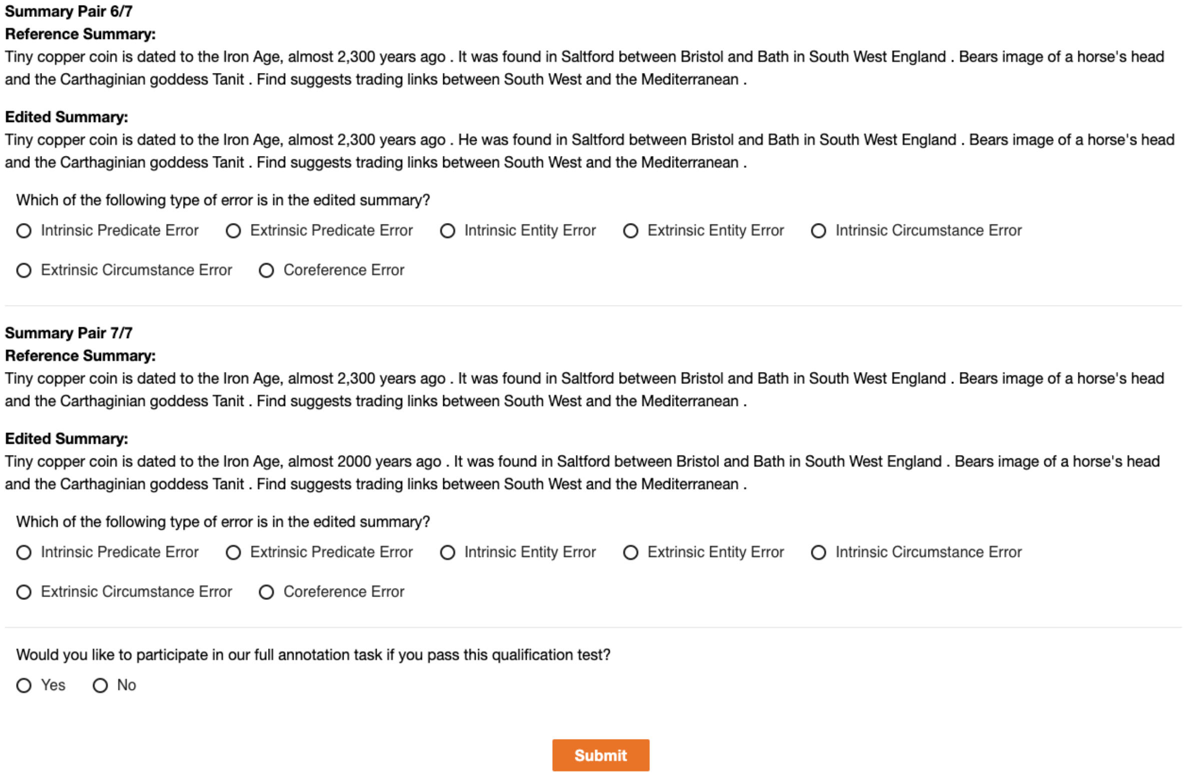
Edited Summary:
Tiny copper coin is dated to the Iron Age, almost 3,200 years ago . It was found in Saltford between Bristol and Bath in South West England . Bears image of a horse's head and the Carthaginian goddess Tanit . Find suggests trading links between South West and the Mediterranean .

Which of the following type of error is in the edited summary?

○ Intrinsic Predicate Error ○ Extrinsic Predicate Error ○ Intrinsic Entity Error ○ Extrinsic Entity Error ○ Intrinsic Circumstance Error

○ Extrinsic Circumstance Error ○ Coreference Error

Figure A3: Screenshot of the qualification task for Task 1 (3/4).



Summary Pair 6/7
Reference Summary:
Tiny copper coin is dated to the Iron Age, almost 2,300 years ago . It was found in Saltford between Bristol and Bath in South West England . Bears image of a horse's head and the Carthaginian goddess Tanit . Find suggests trading links between South West and the Mediterranean .

Edited Summary:
Tiny copper coin is dated to the Iron Age, almost 2,300 years ago . He was found in Saltford between Bristol and Bath in South West England . Bears image of a horse's head and the Carthaginian goddess Tanit . Find suggests trading links between South West and the Mediterranean .

Which of the following type of error is in the edited summary?

○ Intrinsic Predicate Error ○ Extrinsic Predicate Error ○ Intrinsic Entity Error ○ Extrinsic Entity Error ○ Intrinsic Circumstance Error

○ Extrinsic Circumstance Error ○ Coreference Error

Summary Pair 7/7
Reference Summary:
Tiny copper coin is dated to the Iron Age, almost 2,300 years ago . It was found in Saltford between Bristol and Bath in South West England . Bears image of a horse's head and the Carthaginian goddess Tanit . Find suggests trading links between South West and the Mediterranean .

Edited Summary:
Tiny copper coin is dated to the Iron Age, almost 2000 years ago . It was found in Saltford between Bristol and Bath in South West England . Bears image of a horse's head and the Carthaginian goddess Tanit . Find suggests trading links between South West and the Mediterranean .

Which of the following type of error is in the edited summary?

○ Intrinsic Predicate Error ○ Extrinsic Predicate Error ○ Intrinsic Entity Error ○ Extrinsic Entity Error ○ Intrinsic Circumstance Error

○ Extrinsic Circumstance Error ○ Coreference Error

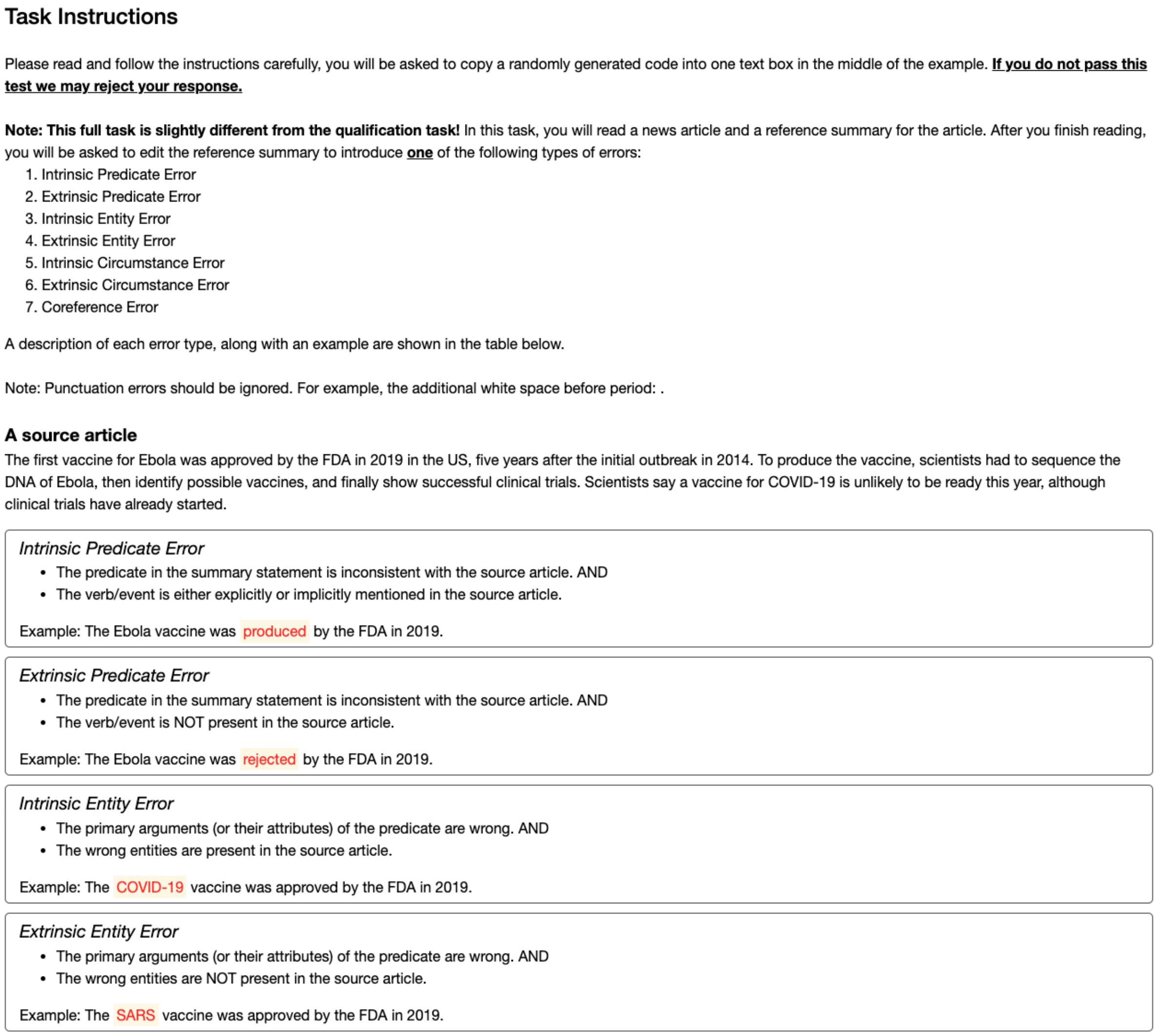
Would you like to participate in our full annotation task if you pass this qualification test?

○ Yes ○ No

Figure A4: Screenshot of the qualification task for Task 1 (4/4).

	Q^2	SummaC	BARTScore	FactCC	BLEURT	BERTScore	QuestEval
QAGS-C	83.5	80.9	80.9	76.4	71.6	69.1	64.2
FRANK	87.8	89.1	86.1	76.4	82.8	84.3	84.0
SummEval	78.8	81.7	73.5	75.9	66.7	77.2	70.1

Table A1: **ROC AUC (%) of faithfulness evaluation metrics in TRUE** (Honovich et al., 2022). All datasets contain CNN/DailyMail. Faithful and unfaithful samples in QAGS-C are balanced; however, in FRANK and SummEval, faithful and unfaithful samples are imbalanced.



Task Instructions

Please read and follow the instructions carefully, you will be asked to copy a randomly generated code into one text box in the middle of the example. **If you do not pass this test we may reject your response.**

Note: This full task is slightly different from the qualification task! In this task, you will read a news article and a reference summary for the article. After you finish reading, you will be asked to edit the reference summary to introduce **one** of the following types of errors:

1. Intrinsic Predicate Error
2. Extrinsic Predicate Error
3. Intrinsic Entity Error
4. Extrinsic Entity Error
5. Intrinsic Circumstance Error
6. Extrinsic Circumstance Error
7. Coreference Error

A description of each error type, along with an example are shown in the table below.

Note: Punctuation errors should be ignored. For example, the additional white space before period: .

A source article

The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.

Intrinsic Predicate Error

- The predicate in the summary statement is inconsistent with the source article. AND
- The verb/event is either explicitly or implicitly mentioned in the source article.

Example: The Ebola vaccine was produced by the FDA in 2019.

Extrinsic Predicate Error

- The predicate in the summary statement is inconsistent with the source article. AND
- The verb/event is NOT present in the source article.

Example: The Ebola vaccine was rejected by the FDA in 2019.

Intrinsic Entity Error

- The primary arguments (or their attributes) of the predicate are wrong. AND
- The wrong entities are present in the source article.

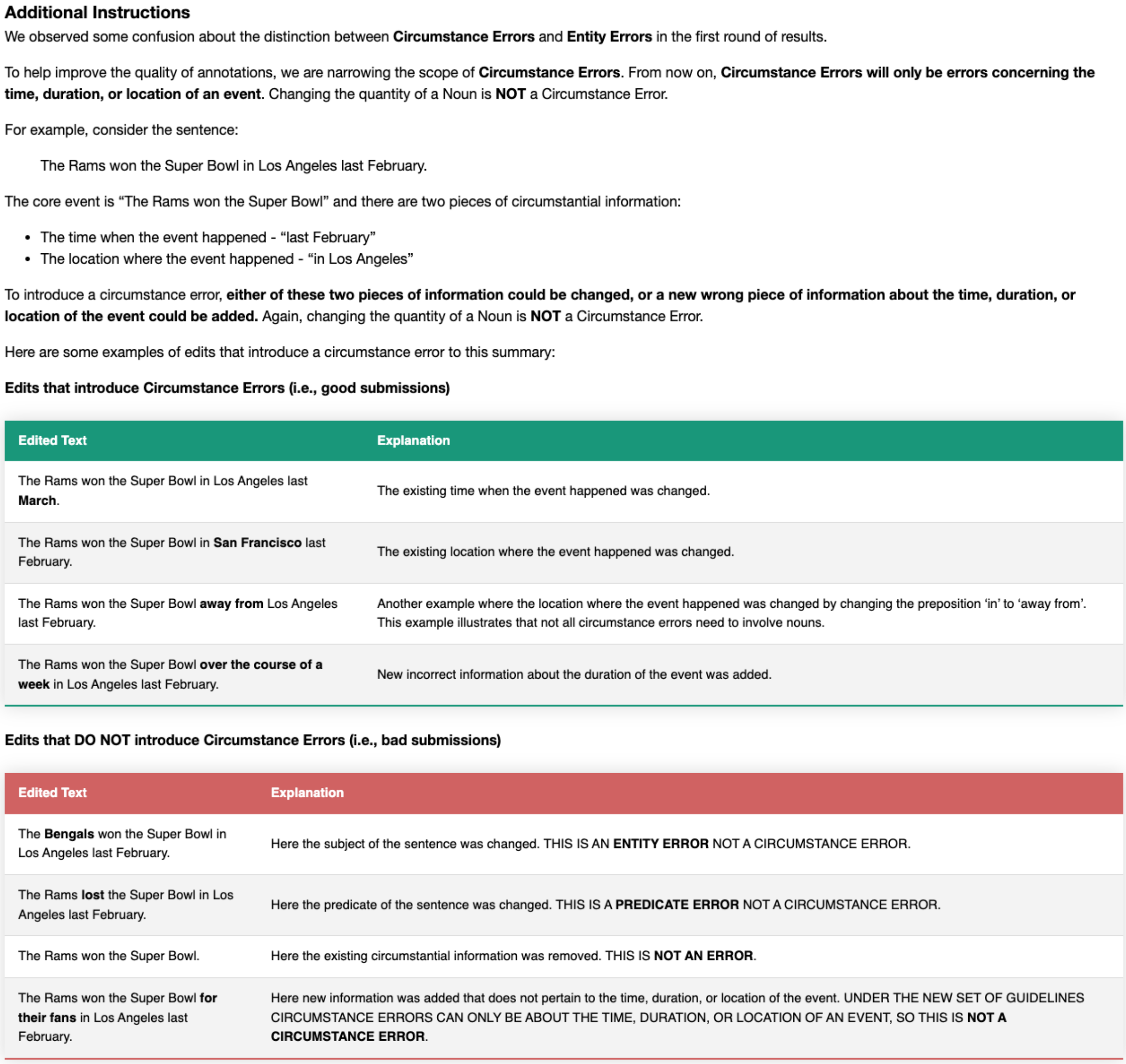
Example: The COVID-19 vaccine was approved by the FDA in 2019.

Extrinsic Entity Error

- The primary arguments (or their attributes) of the predicate are wrong. AND
- The wrong entities are NOT present in the source article.

Example: The SARS vaccine was approved by the FDA in 2019.

Figure A5: Screenshot of the full task for Task 1 (1/3).



Additional Instructions

We observed some confusion about the distinction between **Circumstance Errors** and **Entity Errors** in the first round of results.

To help improve the quality of annotations, we are narrowing the scope of **Circumstance Errors**. From now on, **Circumstance Errors will only be errors concerning the time, duration, or location of an event**. Changing the quantity of a Noun is **NOT** a Circumstance Error.

For example, consider the sentence:

> The Rams won the Super Bowl in Los Angeles last February.

The core event is "The Rams won the Super Bowl" and there are two pieces of circumstantial information:

- The time when the event happened - "last February"
- The location where the event happened - "in Los Angeles"

To introduce a circumstance error, **either of these two pieces of information could be changed, or a new wrong piece of information about the time, duration, or location of the event could be added.** Again, changing the quantity of a Noun is **NOT** a Circumstance Error.

Here are some examples of edits that introduce a circumstance error to this summary:

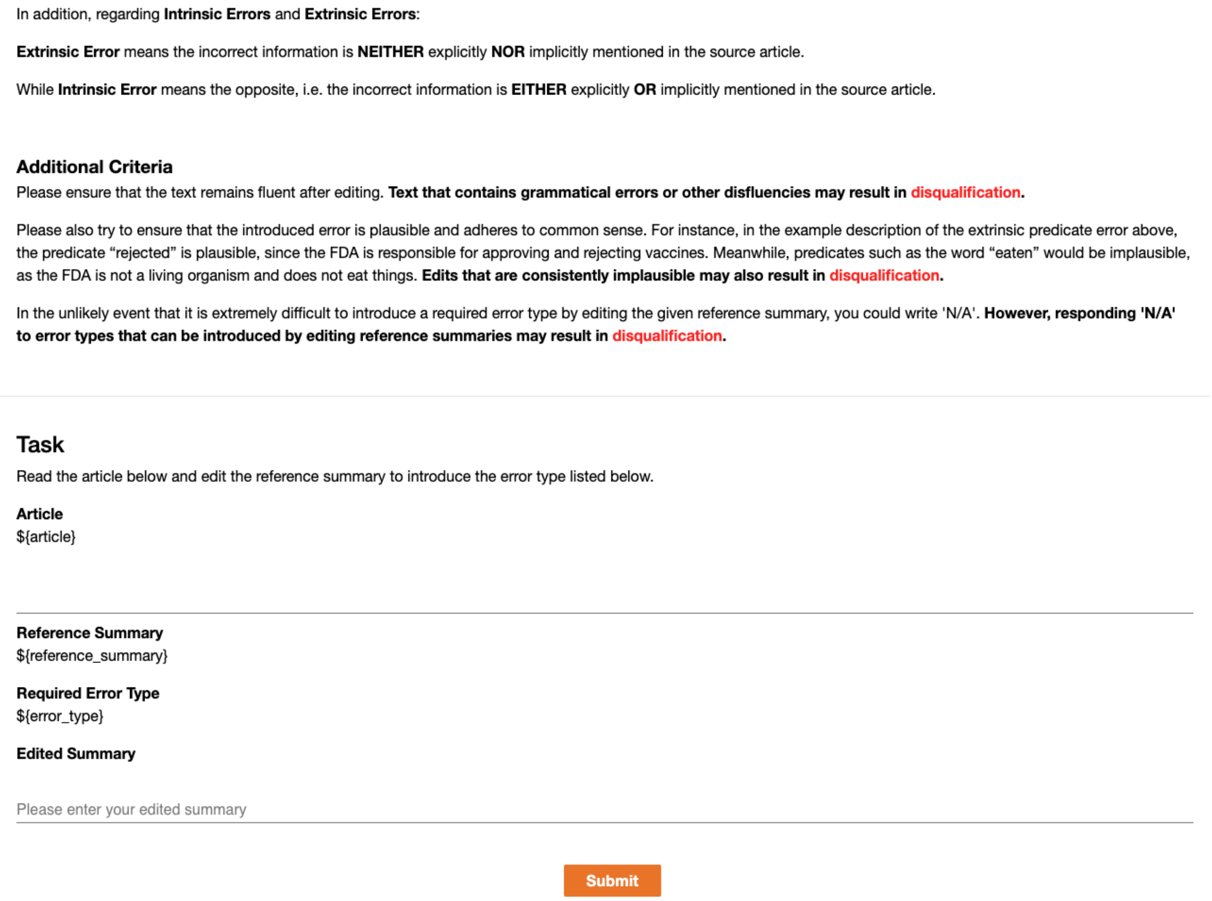
Edits that introduce Circumstance Errors (i.e., good submissions)

Edited Text	Explanation
The Rams won the Super Bowl in Los Angeles last **March**.	The existing time when the event happened was changed.
The Rams won the Super Bowl in **San Francisco** last February.	The existing location where the event happened was changed.
The Rams won the Super Bowl **away from** Los Angeles last February.	Another example where the location where the event happened was changed by changing the preposition 'in' to 'away from'. This example illustrates that not all circumstance errors need to involve nouns.
The Rams won the Super Bowl **over the course of a week** in Los Angeles last February.	New incorrect information about the duration of the event was added.

Edits that DO NOT introduce Circumstance Errors (i.e., bad submissions)

Edited Text	Explanation
The **Bengals** won the Super Bowl in Los Angeles last February.	Here the subject of the sentence was changed. THIS IS AN **ENTITY ERROR** NOT A CIRCUMSTANCE ERROR.
The Rams **lost** the Super Bowl in Los Angeles last February.	Here the predicate of the sentence was changed. THIS IS A **PREDICATE ERROR** NOT A CIRCUMSTANCE ERROR.
The Rams won the Super Bowl.	Here the existing circumstantial information was removed. THIS IS **NOT AN ERROR**.
The Rams won the Super Bowl **for their fans** in Los Angeles last February.	Here new information was added that does not pertain to the time, duration, or location of the event. UNDER THE NEW SET OF GUIDELINES CIRCUMSTANCE ERRORS CAN ONLY BE ABOUT THE TIME, DURATION, OR LOCATION OF AN EVENT, SO THIS IS **NOT A CIRCUMSTANCE ERROR**.

Figure A6: Screenshot of the full task for Task 1 (2/3).



In addition, regarding **Intrinsic Errors** and **Extrinsic Errors**:

Extrinsic Error means the incorrect information is **NEITHER** explicitly **NOR** implicitly mentioned in the source article.

While **Intrinsic Error** means the opposite, i.e. the incorrect information is **EITHER** explicitly **OR** implicitly mentioned in the source article.

Additional Criteria

Please ensure that the text remains fluent after editing. **Text that contains grammatical errors or other disfluencies may result in disqualification.**

Please also try to ensure that the introduced error is plausible and adheres to common sense. For instance, in the example description of the extrinsic predicate error above, the predicate "rejected" is plausible, since the FDA is responsible for approving and rejecting vaccines. Meanwhile, predicates such as the word "eaten" would be implausible, as the FDA is not a living organism and does not eat things. **Edits that are consistently implausible may also result in disqualification.**

In the unlikely event that it is extremely difficult to introduce a required error type by editing the given reference summary, you could write 'N/A'. **However, responding 'N/A' to error types that can be introduced by editing reference summaries may result in disqualification.**

Task

Read the article below and edit the reference summary to introduce the error type listed below.

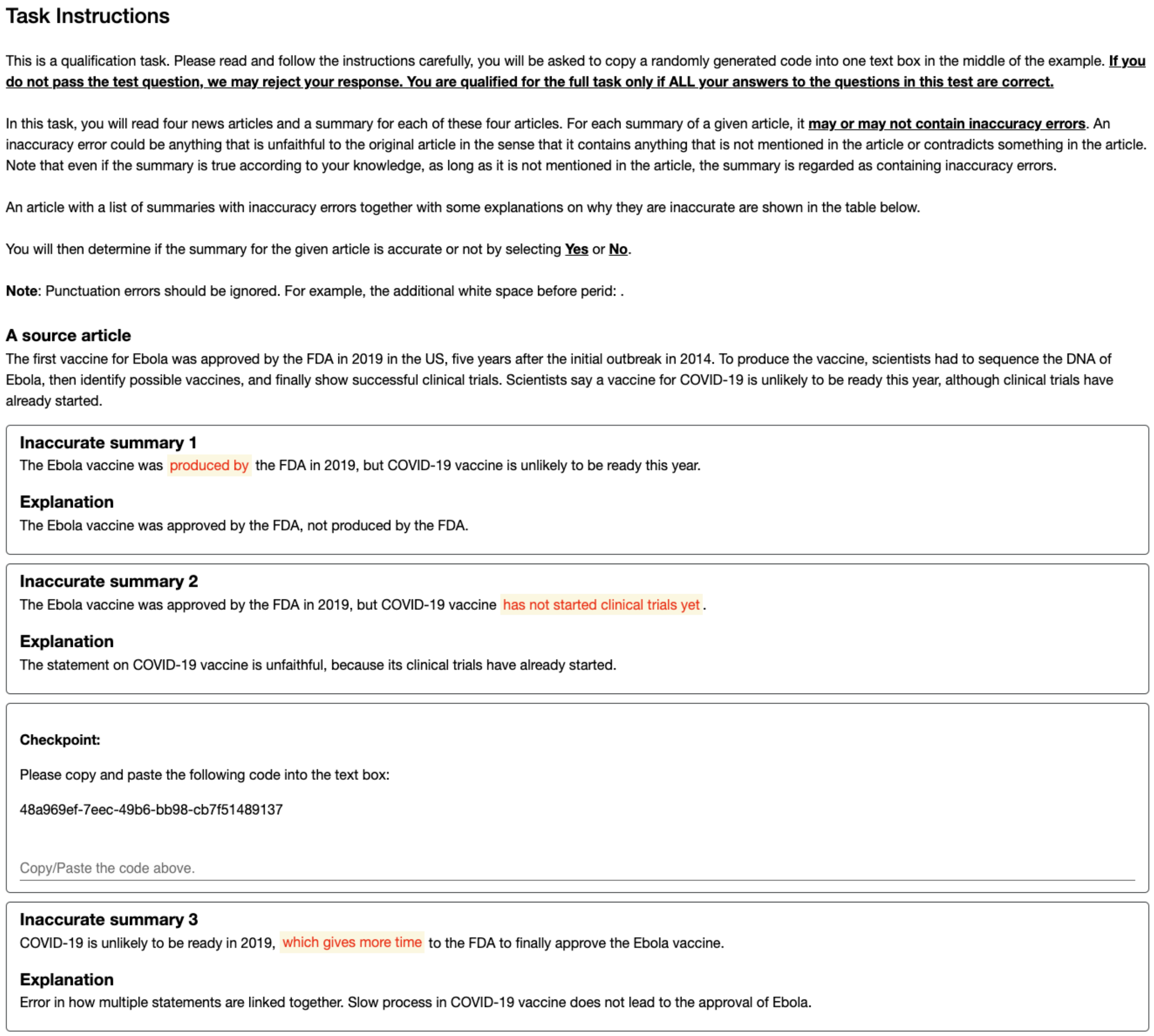
Article
${article}

Reference Summary
${reference_summary}

Required Error Type
${error_type}

Edited Summary

Figure A7: Screenshot of the full task for Task 1 (3/3).



Task Instructions

This is a qualification task. Please read and follow the instructions carefully, you will be asked to copy a randomly generated code into one text box in the middle of the example. **If you do not pass the test question, we may reject your response. You are qualified for the full task only if ALL your answers to the questions in this test are correct.**

In this task, you will read four news articles and a summary for each of these four articles. For each summary of a given article, it **may or may not contain inaccuracy errors**. An inaccuracy error could be anything that is unfaithful to the original article in the sense that it contains anything that is not mentioned in the article or contradicts something in the article. Note that even if the summary is true according to your knowledge, as long as it is not mentioned in the article, the summary is regarded as containing inaccuracy errors.

An article with a list of summaries with inaccuracy errors together with some explanations on why they are inaccurate are shown in the table below.

You will then determine if the summary for the given article is accurate or not by selecting **Yes** or **No**.

Note: Punctuation errors should be ignored. For example, the additional white space before perid: .

A source article

The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.

Inaccurate summary 1

The Ebola vaccine was produced by the FDA in 2019, but COVID-19 vaccine is unlikely to be ready this year.

Explanation

The Ebola vaccine was approved by the FDA, not produced by the FDA.

Inaccurate summary 2

The Ebola vaccine was approved by the FDA in 2019, but COVID-19 vaccine has not started clinical trials yet .

Explanation

The statement on COVID-19 vaccine is unfaithful, because its clinical trials have already started.

Checkpoint:

Please copy and paste the following code into the text box:

48a969ef-7eec-49b6-bb98-cb7f51489137

Copy/Paste the code above.

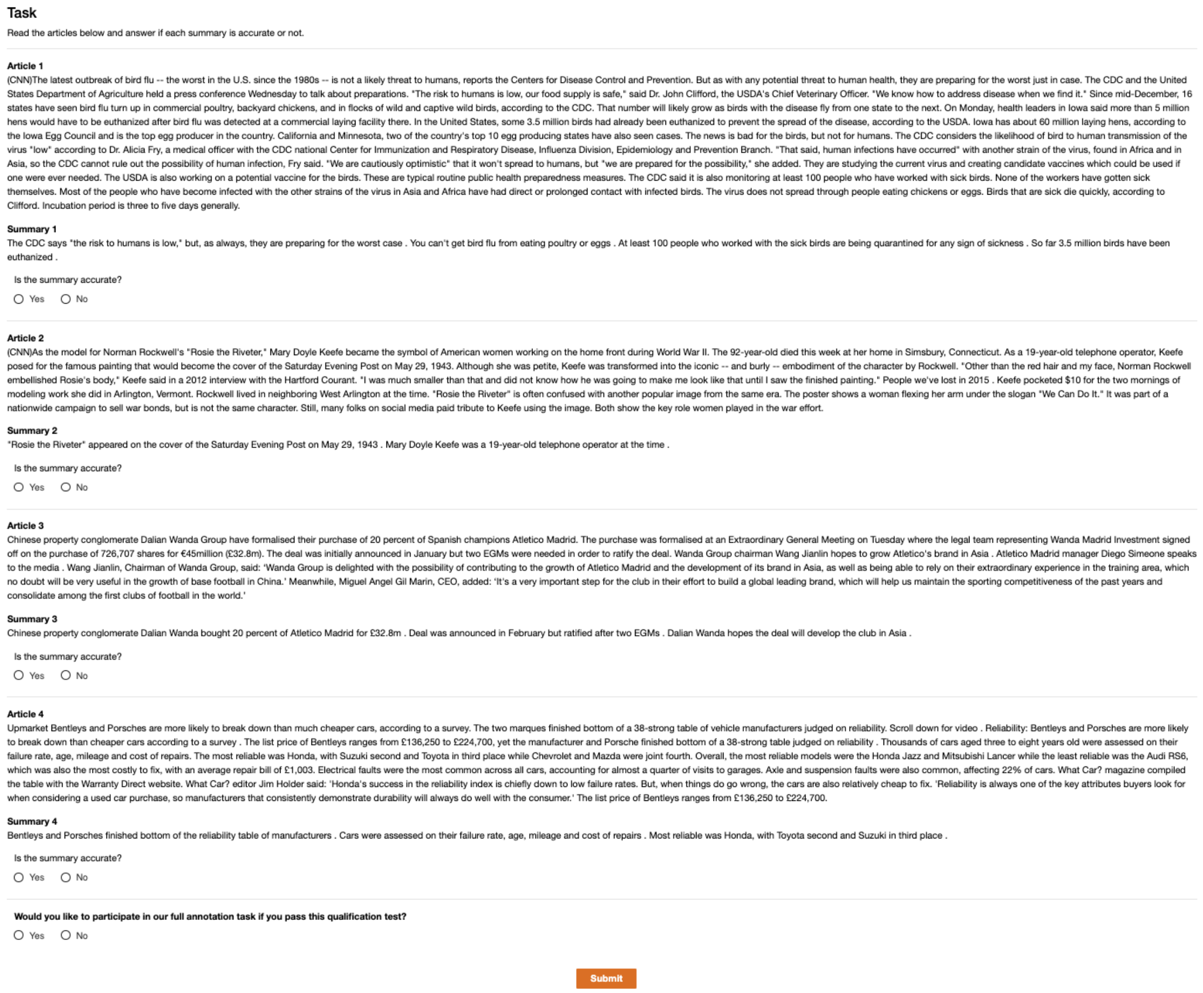
Inaccurate summary 3

COVID-19 is unlikely to be ready in 2019, which gives more time to the FDA to finally approve the Ebola vaccine.

Explanation

Error in how multiple statements are linked together. Slow process in COVID-19 vaccine does not lead to the approval of Ebola.

Figure A8: Screenshot of the qualification task for Task 2 (1/2).



Task

Read the articles below and answer if each summary is accurate or not.

Article 1
(CNN)The latest outbreak of bird flu -- the worst in the U.S. since the 1980s -- is not a likely threat to humans, reports the Centers for Disease Control and Prevention. But as with any potential threat to human health, they are preparing for the worst just in case. The CDC and the United States Department of Agriculture held a press conference Wednesday to talk about preparations. "The risk to humans is low, our food supply is safe," said Dr. John Clifford, the USDA's Chief Veterinary Officer. "We know how to address disease when we find it." Since mid-December, 16 states have seen bird flu turn up in commercial poultry, backyard chickens, and in flocks of wild and captive wild birds, according to the CDC. That number will likely grow as birds with the disease fly from one state to the next. On Monday, health leaders in Iowa said more than 5 million hens would have to be euthanized after bird flu was detected at a commercial laying facility there. In the United States, some 3.5 million birds had already been euthanized to prevent the spread of the disease, according to the USDA. Iowa has about 60 million laying hens, according to the Iowa Egg Council and is the top egg producer in the country. California and Minnesota, two of the country's top 10 egg producing states have also seen cases. The news is bad for the birds, but not for humans. The CDC considers the likelihood of bird to human transmission of the virus "low" according to Dr. Alicia Fry, a medical officer with the CDC national Center for Immunization and Respiratory Disease, Influenza Division, Epidemiology and Prevention Branch. "That said, human infections have occurred" with another strain of the virus, found in Africa and in Asia, so the CDC cannot rule out the possibility of human infection, Fry said. "We are cautiously optimistic" that it won't spread to humans, but "we are prepared for the possibility," she added. They are studying the current virus and creating candidate vaccines which could be used if one were ever needed. The USDA is also working on a potential vaccine for the birds. These are typical routine public health preparedness measures. The CDC said it is also monitoring at least 100 people who have worked with sick birds. None of the workers have gotten sick themselves. Most of the people who have become infected with the other strains of the virus in Asia and Africa have had direct or prolonged contact with infected birds. The virus does not spread through people eating chickens or eggs. Birds that are sick die quickly, according to Clifford. Incubation period is three to five days generally.

Summary 1
The CDC says "the risk to humans is low," but, as always, they are preparing for the worst case . You can't get bird flu from eating poultry or eggs . At least 100 people who worked with the sick birds are being quarantined for any sign of sickness . So far 3.5 million birds have been euthanized .

Is the summary accurate?

○ Yes ○ No

Article 2
(CNN)As the model for Norman Rockwell's "Rosie the Riveter," Mary Doyle Keefe became the symbol of American women working on the home front during World War II. The 92-year-old died this week at her home in Simsbury, Connecticut. As a 19-year-old telephone operator, Keefe posed for the famous painting that would become the cover of the Saturday Evening Post on May 29, 1943. Although she was petite, Keefe was transformed into the iconic -- and burly -- embodiment of the character by Rockwell. "Other than the red hair and my face, Norman Rockwell embellished Rosie's body," Keefe said in a 2012 interview with the Hartford Courant. "I was much smaller than that and did not know how he was going to make me look like that until I saw the finished painting." People we've lost in 2015 . Keefe pocketed $10 for the two mornings of modeling work she did in Arlington, Vermont. Rockwell lived in neighboring West Arlington at the time. "Rosie the Riveter" is often confused with another popular image from the same era. The poster shows a woman flexing her arm under the slogan "We Can Do It." It was part of a nationwide campaign to sell war bonds, but is not the same character. Still, many folks on social media paid tribute to Keefe using the image. Both show the key role women played in the war effort.

Summary 2
"Rosie the Riveter" appeared on the cover of the Saturday Evening Post on May 29, 1943 . Mary Doyle Keefe was a 19-year-old telephone operator at the time .

Is the summary accurate?

○ Yes ○ No

Article 3
Chinese property conglomerate Dalian Wanda Group have formalised their purchase of 20 percent of Spanish champions Atletico Madrid. The purchase was formalised at an Extraordinary General Meeting on Tuesday where the legal team representing Wanda Madrid Investment signed off on the purchase of 726,707 shares for €45million (£32.8m). The deal was initially announced in January but two EGMs were needed in order to ratify the deal. Wanda Group chairman Wang Jianlin hopes to grow Atletico's brand in Asia . Atletico Madrid manager Diego Simeone speaks to the media . Wang Jianlin, Chairman of Wanda Group, said: 'Wanda Group is delighted with the possibility of contributing to the growth of Atletico Madrid and the development of its brand in Asia, as well as being able to rely on their extraordinary experience in the training area, which no doubt will be very useful in the growth of base football in China.' Meanwhile, Miguel Angel Gil Marin, CEO, added: 'It's a very important step for the club in their effort to build a global leading brand, which will help us maintain the sporting competitiveness of the past years and consolidate among the first clubs of football in the world.'

Summary 3
Chinese property conglomerate Dalian Wanda bought 20 percent of Atletico Madrid for £32.8m . Deal was announced in February but ratified after two EGMs . Dalian Wanda hopes the deal will develop the club in Asia .

Is the summary accurate?

○ Yes ○ No

Article 4
Upmarket Bentleys and Porsches are more likely to break down than much cheaper cars, according to a survey. The two marques finished bottom of a 38-strong table of vehicle manufacturers judged on reliability. Scroll down for video . Reliability: Bentleys and Porsches are more likely to break down than cheaper cars according to a survey . The list price of Bentleys ranges from £136,250 to £224,700, yet the manufacturer and Porsche finished bottom of a 38-strong table judged on reliability . Thousands of cars aged three to eight years old were assessed on their failure rate, age, mileage and cost of repairs. The most reliable was Honda, with Suzuki second and Toyota in third place while Chevrolet and Mazda were joint fourth. Overall, the most reliable models were the Honda Jazz and Mitsubishi Lancer while the least reliable was the Audi RS6, which was also the most costly to fix, with an average repair bill of £1,003. Electrical faults were the most common across all cars, accounting for almost a quarter of visits to garages. Axle and suspension faults were also common, affecting 22% of cars. What Car? magazine compiled the table with the Warranty Direct website. What Car? editor Jim Holder said: 'Honda's success in the reliability index is chiefly down to low failure rates. But, when things do go wrong, the cars are also relatively cheap to fix. 'Reliability is always one of the key attributes buyers look for when considering a used car purchase, so manufacturers that consistently demonstrate durability will always do well with the consumer.' The list price of Bentleys ranges from £136,250 to £224,700.

Summary 4
Bentleys and Porsches finished bottom of the reliability table of manufacturers . Cars were assessed on their failure rate, age, mileage and cost of repairs . Most reliable was Honda, with Toyota second and Suzuki in third place .

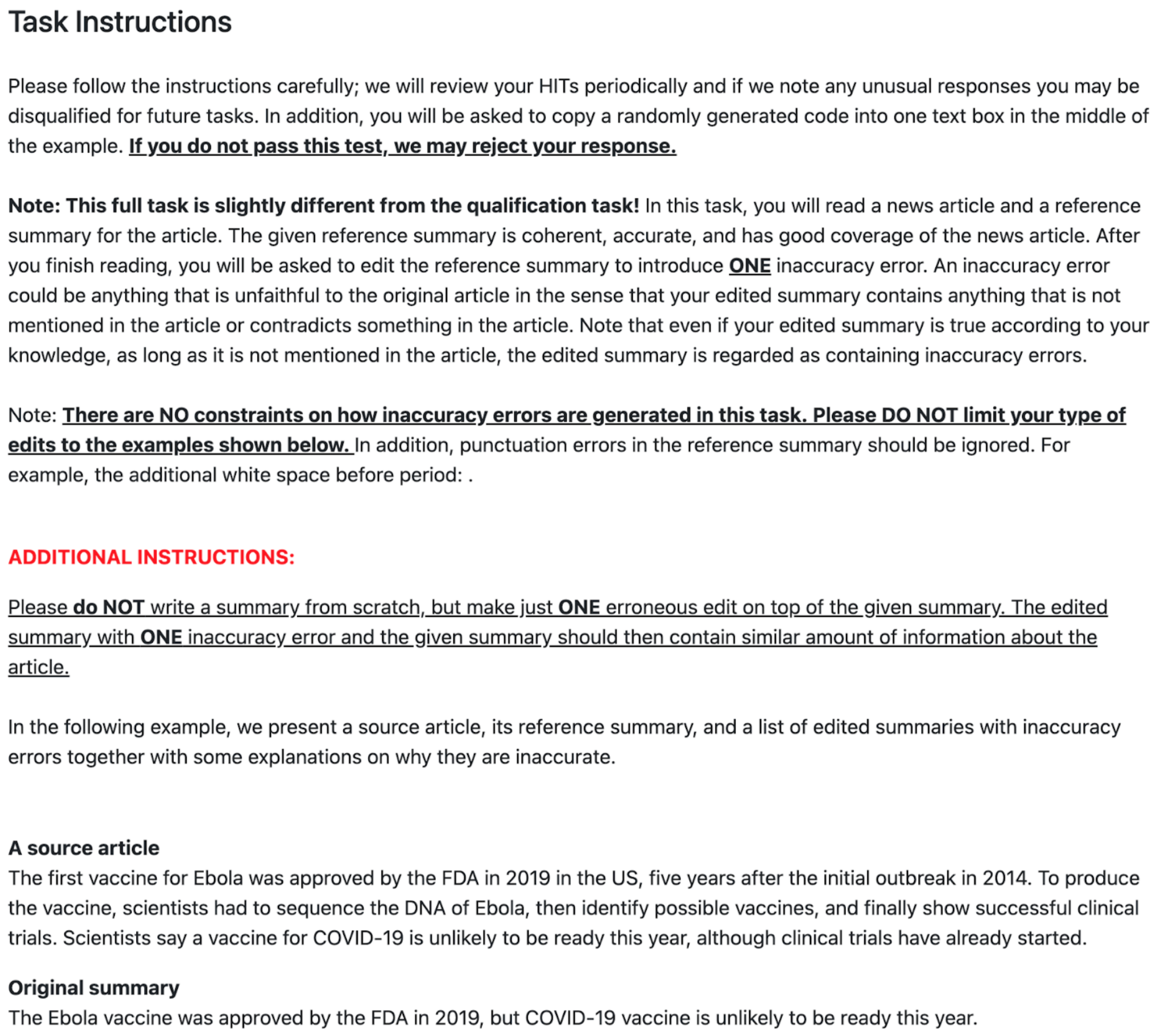
Is the summary accurate?

○ Yes ○ No

Would you like to participate in our full annotation task if you pass this qualification test?

○ Yes ○ No

Figure A9: Screenshot of the qualification task for Task 2 (2/2).



Task Instructions

Please follow the instructions carefully; we will review your HITs periodically and if we note any unusual responses you may be disqualified for future tasks. In addition, you will be asked to copy a randomly generated code into one text box in the middle of the example. **If you do not pass this test, we may reject your response.**

Note: This full task is slightly different from the qualification task! In this task, you will read a news article and a reference summary for the article. The given reference summary is coherent, accurate, and has good coverage of the news article. After you finish reading, you will be asked to edit the reference summary to introduce **ONE** inaccuracy error. An inaccuracy error could be anything that is unfaithful to the original article in the sense that your edited summary contains anything that is not mentioned in the article or contradicts something in the article. Note that even if your edited summary is true according to your knowledge, as long as it is not mentioned in the article, the edited summary is regarded as containing inaccuracy errors.

Note: **There are NO constraints on how inaccuracy errors are generated in this task. Please DO NOT limit your type of edits to the examples shown below.** In addition, punctuation errors in the reference summary should be ignored. For example, the additional white space before period: .

ADDITIONAL INSTRUCTIONS:

Please **do NOT** write a summary from scratch, but make just **ONE** erroneous edit on top of the given summary. The edited summary with **ONE** inaccuracy error and the given summary should then contain similar amount of information about the article.

In the following example, we present a source article, its reference summary, and a list of edited summaries with inaccuracy errors together with some explanations on why they are inaccurate.

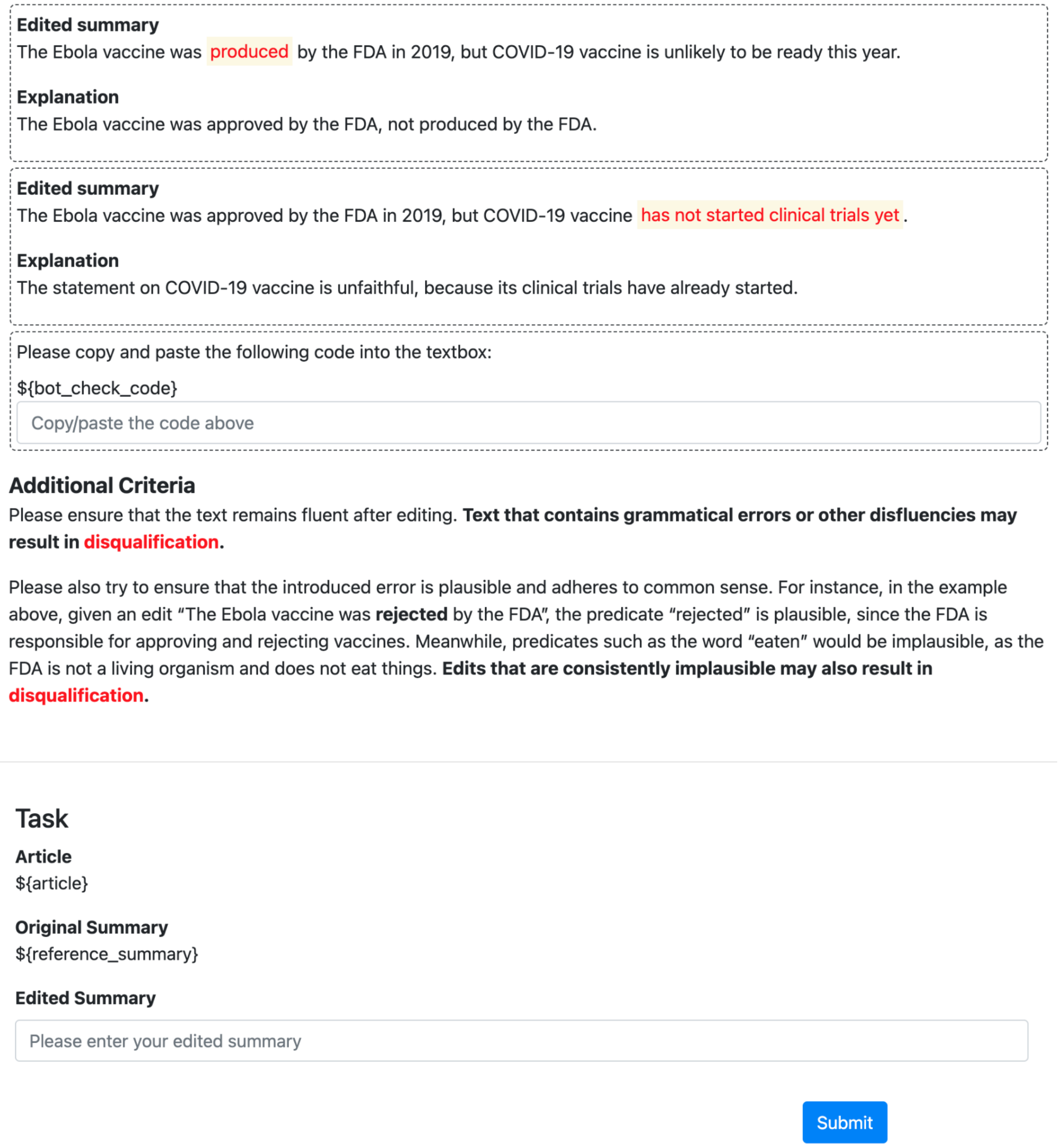
A source article

The first vaccine for Ebola was approved by the FDA in 2019 in the US, five years after the initial outbreak in 2014. To produce the vaccine, scientists had to sequence the DNA of Ebola, then identify possible vaccines, and finally show successful clinical trials. Scientists say a vaccine for COVID-19 is unlikely to be ready this year, although clinical trials have already started.

Original summary

The Ebola vaccine was approved by the FDA in 2019, but COVID-19 vaccine is unlikely to be ready this year.

Figure A10: Screenshot of the full task for Task 2 (1/2).



Edited summary

The Ebola vaccine was produced by the FDA in 2019, but COVID-19 vaccine is unlikely to be ready this year.

Explanation

The Ebola vaccine was approved by the FDA, not produced by the FDA.

Edited summary

The Ebola vaccine was approved by the FDA in 2019, but COVID-19 vaccine has not started clinical trials yet .

Explanation

The statement on COVID-19 vaccine is unfaithful, because its clinical trials have already started.

Please copy and paste the following code into the textbox:

${bot_check_code}

Additional Criteria

Please ensure that the text remains fluent after editing. **Text that contains grammatical errors or other disfluencies may result in disqualification.**

Please also try to ensure that the introduced error is plausible and adheres to common sense. For instance, in the example above, given an edit “The Ebola vaccine was **rejected** by the FDA”, the predicate “rejected” is plausible, since the FDA is responsible for approving and rejecting vaccines. Meanwhile, predicates such as the word “eaten” would be implausible, as the FDA is not a living organism and does not eat things. **Edits that are consistently implausible may also result in disqualification.**

Task

Article

${article}

Original Summary

${reference_summary}

Edited Summary

Figure A11: Screenshot of the full task for Task 2 (2/2).

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the Limitations section

☑ A2. Did you discuss any potential risks of your work?
In the section of Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
CNN/DailyMail is distributed under the MIT license. We plan on releasing our dataset under the MIT license as well pending legal approval. The LICENSE will be provided alongside the dataset as a text file on GitHub when the paper is published.
UPDATE AFTER REVIEW: We recently learned that the data we were using is distributed under an Apache 2.0 license instead of MIT.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
CNN/DailyMail is distributed under the MIT license. Our BUMP dataset is derived from CNN/DailyMail. We plan on releasing BUMP under the MIT license as well pending legal approval. The LICENSE will be provided alongside the dataset as a text file on GitHub when the paper is published. Therefore, the data collection and distribution of BUMP is consistent with the license in CNN/DailyMail.
UPDATE AFTER REVIEW: We recently learned that the data we were using is distributed under an Apache 2.0 license instead of MIT.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We manually checked all collected human annotations; see Section 3.2 and 3.3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3, and Appendix A and B

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Table 2 and its descriptions in Section 3.2 and 3.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4 and 5

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3, Appendix A and B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section of Ethics Statement, and Appendix A.1 and B.1

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
The BUMP dataset is derived from CNN/DM. Attribution is provided in Section 3.1. Since CNN/DM is distributed under the MIT license (which allows modifications), we did not discuss with annotators on how their modified summaries will be used.
UPDATE AFTER REVIEW: We recently learned that the data we were using is distributed under an Apache 2.0 license instead of MIT.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
We used a review process internal to our organization with HCI research scientists.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix A.1 and B.1